格致方法
定量研究系列

·革新研究理念
·丰富研究工具
·最权威、最前沿的定量研究方法指南

吴晓刚 / 主编
【美】保尔·D. 埃里森(Paul D. Allison) / 等著

高级回归分析

格致出版社 上海人民出版社

世纪出版

格致方法

定量研究系列

高级回归分析

本书由5本讨论高级回归分析的小册子组成，分别是《固定效应回归模型》、《现代稳健回归方法》、《删截、选择性样本及截断数据回归模型》、《分位数回归模型》及《空间回归模型》。《固定效应回归模型》介绍了多种形式的固定效应回归模型，讨论了如何在固定效应模型及随机效应模型之间作出选择；《现代稳健回归方法》通过一套统一的符号系统，介绍了不同来源的多种稳健回归方法，以及它们彼此之间的联系；《删截、选择性样本及截断数据回归模型》是有关删截数据、选择性样本数据及截断数据的最新研究；《分位数回归模型》提出了分位数和分位数函数的概念，阐述了分位数回归模型，讨论了它们的估计和推断方法，并通过具体的例子演示了对分位数回归估计值的解释；《空间回归模型》介绍了两种应用最广泛的空间回归模型：空间定距因变量和空间性误差模型。

上架建议：社科研究方法

ISBN 978-7-5432-1899-4

9 787543 218994 >

定价：78.00元

易文网：www.ewen.cc

格致网：www.hibooks.cn

格致方法
定量研究系列

吴晓刚 / 主编

【美】保尔 · D. 埃里森(Paul D. Allison) / 等著

高级回归分析

SAGE Publications ,Inc.

格致出版社 上海人民出版社

图书在版编目(CIP)数据

高级回归分析/(美)埃里森(Allison, P. D.)等著;吴晓刚主编. —上海:格致出版社:上海人民出版社,2011

(格致方法·定量研究系列)

ISBN 978-7-5432-1899-4

Ⅰ. ①高… Ⅱ. ①埃… ②吴… Ⅲ. ①回归分析-研究 Ⅳ. ①O212.1

中国版本图书馆CIP数据核字(2010)第254745号

责任编辑 罗 康

封面装帧 人马艺术工作室·储平

格致方法·定量研究系列

高级回归分析

[美]保尔·D.埃里森 等著

吴晓刚 主编

出 版 世纪出版集团 www.ewen.cc 格致出版社 www.hibooks.cn 上海人民出版社

(200001 上海福建中路193号24层)

编辑部热线 021-63914988

市场部热线 021-63914081

发 行 世纪出版集团发行中心

印 刷 上海图宇印刷有限公司

开 本 787×1092毫米 1/16

印 张 34.75

插 页 1

字 数 504,000

版 次 2011年8月第1版

印 次 2011年8月第1次印刷

ISBN 978-7-5432-1899-4/C·44

定 价 78.00元

出版说明

《高级回归分析》是“格致方法·定量研究系列”之一种，由 5 本讨论高级回归分析的小册子组成，分别是《固定效应回归模型》、《现代稳健回归方法》、《删截、选择性样本及截断数据回归模型》、《分位数回归模型》及《空间回归模型》。

《固定效应回归模型》介绍了多种形式的固定效应回归模型，讨论了如何在固定效应模型及随机效应模型之间作出选择；《现代稳健回归方法》通过一套统一的符号系统，介绍了不同来源的多种稳健回归方法，以及它们彼此之间的联系；《删截、选择性样本及截断数据回归模型》是一本有关删截数据、选择性样本数据及截断数据的最新研究；《分位数回归模型》提出了分位数和分位数函数的概念，阐述了分位数回归模型，讨论了它们的估计和推断方法，并通过具体的例子演示了对分位数回归估计值的解释；《空间回归模型》介绍了两种应用最广泛的空间回归模型：空间定距因变量和空间性误差模型。

总序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996 年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫"调查数据分析"，第三门叫"研究设计"。除此以外，还有如"定类数据分析"、"人口学方法与技术"、"事件史分析"、"多层线性模型"等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院（Institute for Social Research）定量社会研究方法项目的一些讨论会，受益良多。

2003年，我赴港工作，在香港科技大学社会科学部，教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课"社会科学里的统计学"(Statistics for Social Science)为所有研究型硕士生和博士生的必修课，而第二门课"社会科学中的定量分析"为博士生的必修课(事实上，大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的"消费者"而非"生产者"。作为"消费者"，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。"社会科学里的统计学"(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。"社会科学中的定量分析"则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的"定量社会科学研究"丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选

一批翻译,以飨中文读者时,我非常支持这个想法,因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力,还要有对实质内容有较深的理解能力,而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容,只有语言能力是远远不能胜任的。在短短的一年时间里,我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程,他们目前大部分是香港科技大学的硕士和博士研究生,受过严格的社会科学统计方法的训练,也有来自美国等地对定量研究感兴趣的博士研究生。他们是:

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛,硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊,应用社会经济研究中心研究员李俊秀;香港大学教育学院博士研究生洪岩璧;北京大学社会学系博士研究生李丁、赵亮员;中国人民大学人口学系讲师巫锡炜;中国台湾"中央"研究院社会学所助理研究员林宗弘;南京师范大学心理学系副教授陈陈;美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛;美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景,书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异,校对也未免挂一漏万,术语的标准译法方面还有很大的改进空间,但所有的参与者都做了最大的努力,在繁忙的学习和研究之余,在不到一年的时间内,完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外,还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林,协助我通读了全稿,在此我也致以诚挚的谢意。有些作者,如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授,也参与了审校工作。

由于所选每本书都有一篇序言,对相关方法的背景和应用作了很好的介绍,我们均予以保留,内容在此不再赘述。为了方便起见,我们将内容相似的书目集册出版,每册三至五本不等,共八册,它们分别是:《线性回归分析基础》、

《高级回归分析》、《广义线性模型》、《列表数据分析》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》和《数据分析方法五种》。所冠书名未必能精准涵盖其中的内容，读者可自行参阅每本书的序言或目录。

我们希望本丛书的出版，能为推动国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

目 录

固定效应回归模型

序 3
第 1 章　绪言 5
第 2 章　线性固定效应模型:基本原理 10
第 3 章　固定效应 Logistic 回归 31
第 4 章　计数变量的固定效应模型 50
第 5 章　事件史数据的固定效应模型 67
第 6 章　固定效应结构方程模型 84
附录 1　第 2 章到第 5 章例题的 Stata 程序 95
附录 2　第 6 章例题的 Mplus 程序 108
注释 114
参考文献 116
译名对照表 120

现代稳健回归方法

序 125
第 1 章　绪言 127

第 2 章　重要背景　134
第 3 章　稳健性、抗扰性与最小二乘回归　157
第 4 章　线性模型的稳健回归　174
第 5 章　稳健回归的标准误　195
第 6 章　广义线性模型中的权势案例　203
第 7 章　结论　214
附录　稳健回归的软件选择　217
注释　219
参考文献　221
译名对照表　227

删截、选择性样本及截断数据的回归模型

序　231
第 1 章　概论　233
第 2 章　删截数据的 Tobit 模型　242
第 3 章　选择性样本模型和截断回归模型　259
第 4 章　基本模型的扩展　272

第 5 章　应注意的问题 280
附录 1　截断正态分布变量的期望值 293
附录 2　切希尔和艾利时(Chesher & Irish)的正态性及异方差检验 295
注释 297
参考文献 298
译名对照表 302

分位数回归模型

序 305
第 1 章　引言 306
第 2 章　分位数和分位数函数 312
第 2 章　附录 327
第 3 章　分位数回归模型及其估计量 329
第 4 章　分位数回归的推论 349
第 5 章　分位数回归估计值的解释 361
第 6 章　单调转换 QRM 的解释 381
第 7 章　实例:1991 年和 2001 年的收入不平等 395
附录　STATA 命令 417
注释 434
参考文献 435
译名对照表 438

空间回归模型

序 443
前言 444
第 1 章 导论 446
第 2 章 空间滞后因变量 479
第 3 章 空间误差模型 508
第 4 章 扩展 519
附录 软件选项 529
注释 532
参考文献 533
译名对照表 539

固定效应回归模型

作 者 简 介

保尔 · D. 埃里森

保尔 · D. 埃里森(Paul D. Allison)是宾夕法尼亚大学社会学教授，讲授的研究生高级课程包括事件史分析、分类数据分析以及能够处理潜变量的结构方程模型。他有 7 本专著，并发表了 50 多篇论文。每年暑假，他都会召开一个为期 5 天的工作坊，向来自全美的约 100 名研究者讲授生存分析和 Logistic 回归分析。曾经荣获古根海姆学者奖(Guggenheim Fellow)，并于 2001 年获得拉扎斯菲尔德奖(Lazarsfeld Award)，以奖励他在社会学研究方法领域的突出贡献。

译 者 简 介

李丁

李丁，北京大学社会学系本科和硕士毕业，现为北京大学社会学系博士研究生。

序

在最近的一次会议上，我聆听了某研究者分析国家一年度(country-year)数据的报告，理应使用固定效应模型，他用的却是随机效应模型。而那篇文章却受到了来自不同社会科学背景的学者的热烈欢迎。显然，在诸多社会科学专业里，就如何选用固定效应模型和随机效应模型还存在很多疑惑，很多人甚至还不清楚这些模型有何用处。无疑，埃里森讨论的是这两种模型更重要和一般的方面。本书将很好地满足"社会科学定量研究方法丛书"在这一主题上的需要，尤其是考虑到现在获得跟踪调查数据(panel data)越来越容易的现实①。

上述国家一年度数据代表着这样一种数据类型，在这种数据中，个体案例得到了历时的(多次)观察。动态跟踪调查(panel survey)之所以近年来非常流行，一个重要的原因是跟踪数据允许研究者把握社会的发展变化，而把握这种变化是真正理解社会机制的必要条件。尽管有些跟踪调查每年都会观测一次，例如英国家户跟踪调查(British Household Panel Study Survey，简称 BHPS)开始于 1991 年，目前仍在持续进行；其他一些则只有少数几轮调查，例如全美青少年健康跟踪调查(National Longitudinal Study of

---

① Panel Data 在经济学文献中通常被翻译为面板数据(如《面板数据计量经济学》[*Panel Data Econometrics*]，曼纽尔·阿雷拉诺著，朱平芳、徐伟民译，上海财经大学出版社，2008 年 10 月)或综列数据(如《计量经济学：现代观点》，[美]J. M. 伍德里奇著，费剑平译，中国人民大学出版社，2003 年 3 月)。面板数据分析(Panel Data Analysis)与时间序列分析(Time Series Analysis)及横截面数据回归分析(Regression Analysis with Cross-Sectional Data)构成计量经济学的三大内容。其中横截面数据是一个时点上收集的不同观察对象的数据，比方说，就一次人口普查来说就是一个截面研究；时间序列数据通常是一个观察单位在不同时点的观察结果构成的数据，如我国 1978 年以来，每一年的 GDP 的增长速度数据就构成一个时间序列数据。而 Panel Data 将这两种数据的特性综合在一起，首先在同一时点上对不同的案例(通常为总体中的一个规模相对较小的子样本)的多个特征进行了观测；其次，对每一个案例在不同时点进行了多次观测；由此所得到的数据就是 Panel Data。面板数据这一翻译尽管因计量经济学而流传甚广，但"面板"的中文字义与英文 panel 含义相差甚远。Panel Survey 翻译成为面板调查显得笨拙，可翻译为小样本重复调查、固定样本长期追踪调查、追踪调查、纵贯调查或者历时调查等等。其中追踪调查或跟踪调查即带有在一段时间内对固定样本进行多次调查的含义。此外，根据艾尔·巴比的观点，Longitudinal Data 为历时研究，它包括趋势研究、队列研究和追踪研究三类。——译者注

Adolescent Health in the United States)只在 1994 年到 2002 年间进行了 3 轮调查。

不管分析单位是个人、单位还是国家，回归模型中每个案例在不同时点上的残差都将存在一定的相关或互相依赖，这通常是因为不同案例在某些未被观察到的特征上存在差异造成的。此时，回归模型有关误差项相互独立的假定被违背(尽管这个一般规律同样适应于限值因变量[limited dependent variable]回归，但这里我们将讨论限定在线性模型上)。

固定效应模型和随机效应模型都能解决残差相关问题。但固定效应模型做得彻底得多。用埃里森的话说，这些模型“将每个个体作为其自身的控制。”经此处理，它们实际上就控制了所有稳定的、未被观测到的变量，就像这些变量实际得到了观测并被纳入模型一样。就此而言，这些模型所起的作用和实验设计中的随机分配如出一辙。

本书作者在过去 30 年间为社会科学研究方法作出了持续的贡献，涉及诸多重要的主题。他撰写的《事件史分析》(Event History Analysis, 1984)，至今仍是社会科学领域介绍事件史数据分析著作的榜样和标准。确切地说，在本书中，埃里森介绍了多种形式的固定效应模型——可以用于连续因变量的、分类因变量的、计数因变量的甚至结构方程情境等——并且讨论了如何在固定效应模型及随机效应模型之间作出选择，这一讨论对于本序言开始时提及的那位报告人将大有裨益。

廖福挺(Tim Futing Liao)

第1章 | 绪言

多年以来，统计学领域最具挑战性的议题，是如何创造一些方法以从非实验数据中进行有效的因果推论。而在这一议题内最难的问题，是如何从统计上控制无法观测的变量。对于实验主义者而言，问题的解决方案非常简单：随机分配(random assignment)。通过将研究对象随机分配到实验组(treatment group)，可使这些小组在研究对象各属性上几乎相似，不管这些属性是可观测的还是不可观测的。但是在非实验研究中，控制这些潜在干扰变量的传统办法就是测量它们，并把它们放到回归模型里。没有测量就没有控制。

在本书中，我描述了一些被称为固定效应模型的回归模型，这些模型使得我们有可能对那些没有或无法被测量的变量进行控制。基本的思想非常简单：用每个个体作为其自身的控制(因素)。例如，如果你想弄清婚姻是否能减少惯犯们(chronic offenders)的再犯行为(recidivism)，可以通过对个体结婚前后遭拘捕的比率进行比较。假定其他情况都不变(这是一个很大的假定)，前后两个时期拘捕率的差异可以作为婚姻对该个体产生的效果的估计。如果我们将人群中不同个体的这一差异进行平均，就能得到"平均处置效应"(average treatment effects)的估计值。这一估计控制了惯犯们所有的稳定属性。它同时控制了容易被测量的变量，诸如性别、民族、种族、出生地，以及更难被控制的变量，如智商、儿童期父母的照料情况、遗传结构等。虽然它不控制诸如就业状况、收入之类的时变变量，但这些变量通过常规的办法——对其进行测量并放入回归模型——就可以得到控制。

再举一个例子，假如你想研究打电脑游戏的时间是否会影响小孩的学习成

绩。你在几个时点上对样本里的小孩都测量了这两个变量。针对每个小孩，都用打电脑游戏的时间对其学习成绩估计一个回归，然后将得到的回归参数进行平均，就可以完成一个基础的固定效应模型。因为只有小孩自身(within-child)的变化被用来估计回归参数，小孩的所有固定属性都得到了控制。

使用固定效应模型有两个基本的数据要求：第一，对于每一个个体，因变量至少要被测量两次。这些测量结果应该具有直接的可比性，也就是说，它们具有同样的意义和度量单位。第二，样本中应该有相当比例的案例的关键自变量在不同时点上的取值有所变化。固定效应模型在估计诸如性别、民族之类的非时变变量的作用效果上几乎毫无用处。当然，有些统计学家认为谈论这些变量的因果效应根本就毫无意义(Sobel, 1995)。

为什么非得用一本书来介绍固定效应模型呢？第一，不同类型的因变量需要使用不同的方法，不管是定距的、定性的、计数的因变量，还是事件时间。第二，对于特定类型的因变量，通常有两种及以上的方法来使用固定效应模型，我们需要理解它们的异同。第三，也是最具挑战性的是，若被测量的自变量并非"严格的外生(exogenous)变量"——例如，因变量在某个时点的取值会影响此后时点上自变量的取值——还需要求助于一些特殊办法(而且并非总能找到此类办法)。

"固定效应模型"这一概念经常与"随机效应模型"形成对照。很可惜，这一术语是众多误解和疑惑的起源。以传统的观点来看，固定效应模型将个体间未被观察的差异作为一套固定的参数，它们要么可以被直接估计出来，要么可以在估计方程中被抵消掉(partial out)。而在随机效应模型中，未被观察的差异被处理成为具有特定概率分布的随机变量。

如果求助于有关实验设计的文献以解释这一差异，你会找到如下论述：

> 如果实验所用的处置水平(treatment level)恰好是推论所试图……的某几个水平时，通常把处置效应看作是固定的。如果试图推论的处置效应的范围比实验中所用的要大，或者处置水平并未经过有目的的选择……通常的做法是把处置水平看做是随机的(LaMotte, 1983:138—139)。

然而，这种描述在非实验情境下是无益的，因为根据他们的建议，几乎在任何情境下随机效应模型都更为可取。没有什么比这更荒谬的了。

从更新潮的框架来看（Wooldrige，2002），未被观测的差异通常被当作随机变量。这时，将这两种模型区别开来的是已观测的变量与未被观测的变量之间的相关结构。在随机效应模型中，未被观测的变量被假定与所有观测变量之间不相关（或者，更严格地说，在统计上互相独立）。而在固定效应模型中，允许未被观测的变量与任何已观测的变量之间存在相关（这实际上相当于将未被观测的变量当作固定的参数）。除非你允许这样的相关，否则就不能真正控制这些未被观测变量的作用。正因如此，固定效应模型才有吸引力。

当然，固定效应模型也有一些潜在的严重不足。前面已经提到，传统的固定效应模型不能对不随时间变化的变量产生任何估计。在本书中，我们将看到一些用来估计此类变量（如性别和民族）的效果的办法，但是这些估计并没有真正控制不可观测的变量。

第二，在很多情况下，固定效应估计产生的标准误要比随机效应估计的大得多，从而导致更大的 P 值和更宽的置信区间。原因很简单，随机效应模型既使用了个体内信息，又使用了个体间信息，而固定效应估计只使用了个体内信息，在根本上忽略了个体间差异的信息。如果自变量取值在个体间存在很大差异，而在同一个体不同时点上的变化不大，那么固定效应估计将很不精确。

例如，用固定效应模型估计教育（受教育年数）对工资收入的影响就很困难。尽管受教育年限会有一定的变化，但绝大多数人是在完成了学业后才开始有工资的。一部分人进入劳动力市场后会争取继续教育，但相对于个体间的差异，同一个体不同时点的教育差异要小得多。另外，那些成年后教育水平仍然有所变化的人可能根本就不同于那些教育水平保持不变的人。

那我们为什么要抛弃个体间的变异呢？这是因为这些变异很可能与个体未被观测的属性混在一块。固定效应模型的思路是避免使用这些“被污染”的变异，而只使用那些能够对我们感兴趣的参数产生近似无偏估计的变异。用统计学的话说，我们牺牲了效率以减小偏差。在非实验研究（nonexperimental studies）中，我想这通常是一个值得的交换。不过，必须记住的是，固定效应模

型无法控制随着时间而发生变化的未被观测的变量。例如，在探索婚姻对累犯的影响的研究中，结婚很有可能与收入的增加联系在一起。因此，除非收入变量被明确地纳入回归模型，否则，估计得到的婚姻的影响实际上将代表收入的影响。

有意思的是，固定效应方法经常被用在随机化的实验中，以提高效率（例如，减小抽样变异程度）而非减小偏误。在交叉设计中（Senn，1993），每个研究对象会在不同的时点接受两次及以上的不同的实验处理（treatment），这些不同处理出现的先后顺序是随机选定的。因此，这些实验处理应该与实验对象之间未被观测到的差异不存在实质关联。此外，根据设计，自变量（实验处理）的所有变化都是个体内的，而不是个体间的，因此，忽略研究对象之间的差异并不会损失任何信息。实际上，因为没有将个体间的变异作为误差项的一部分，固定效应分析很可能产生理想的低标准误。

固定效应方法的另一个诱人之处在于，实现这些方法的软件已经随处可得。例如，对于第 2 章的基本线性模型，一般最小二乘法回归软件就够用了。而第 6 章的高级线性模型可以通过很多用来做结构方程模型的程序进行估计。第 3 章的 logstic 回归模型，如果是两期数据，常规的 logistic 程序就够用了。如果是多期数据，则可以用条件 logit 程序解决，这种程序在绝大多数综合性的统计软件包中都有。用于计数数据的固定效应模型（第 4 章）可以用常规的泊松或负二项回归软件进行估计。最后，第 5 章的事件史模型可以用 Cox 回归标准程序或常规 logit 程序（在事件不重复发生的情况下）进行估计。

要想从本书得到最大的益处，你应该已经对基本的统计推论原则有所了解，包括标准误、置信区间、假设检验、p 值、偏差、有效性等等（Lewis-Beck，1995）。对于具体的章节，你应该对作为固定效应方法基础的那些特定回归方法有所了解。这些方法包括第 2 章的一般线性回归（Allison，1999b）、第 3 章的 logistic 回归（Allison，1999a；Pampel，2000）、第 4 章的泊松及负二项回归（Dunteman & Ho，2005）、第 5 章的 Cox 回归（Allison，1984）以及第 6 章的线性结构方程模型（Long，1983）。

第 2 章到第 5 章例题的运算我用的是 Stata 软件（www. stata. com），它有

大量用于固定效应回归的命令。这些章节所有例题的 Stata 命令在附录 1 中都可以找到。第 6 章使用的是 Mplus 软件(www. statmodel. com),这一章的命令见附录 2。本书的部分内容来自于笔者此前在 SAS ®出版社出版的《用 SAS 对纵贯数据进行固定效应回归分析》一书(*Fixed Effects Regression Methods for Longitudinal Data using SAS* ®, 2005, SAS Institue Inc.)。希望了解如何用 SAS 进行固定效应分析的读者可以参看该书。

第 2 章 | 线性固定效应模型:基本原理

在这一章,我们讨论的固定效应模型,要求数据中的因变量为定距测量变量,并且因变量与自变量之间是线性决定关系。在数据中,我们有一组个体($i = 1, \cdots, n$),并且每一个体都至少在两个时点($t = 1, \cdots, T$)上得到测量。在这里,每一时点常被称为一个"时期"。

下面是模型的表示方法:我们令 y_{it} 表示因变量,用向量x_{it}表示一套在不同时点有所变化的自变量,另外还有一套不随时间变化的自变量z_{it}(如果你觉得使用向量不舒服,可以把它们当成单变量来解释)。我们为 y 建立的基本模型如下:

$$y_{it} = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i + \varepsilon_{it} \qquad [2.1]$$

其中,μ_t 是截距,每一个时期都可以不同;β 和 γ 是系数向量。尽管方程 2.1 看起来像是严格的截面数据模型(strictly cross-sectional),但x_{it}向量要纳入时滞 x 变量(lagged versions of x),一点障碍也没有,只不过要求研究者必须至少有三期数据,才能估计一个一期时滞模型(a model with a lag of one period)。

在上式中,两个"误差"项 α_i 和ε_{it}的特性表现彼此相同。每一个体在每个不同时点都有一个不同的 ε_{it};但 α_i 只在不同个体之间有所不同,不随时间变化而变化。这样,我们可以认为 α_i 代表着所有未被观测到的非时变变量对于 y 的综合影响。相反,ε_{it}代表每一时点上的纯粹随机变动。

现在,我要对 ε_{it} 做一个很强的假定,即每个 ε_{it} 的均值为 0,方差不变(对所有的 i 和 t),并且在统计上独立于所有其他因素(y 除外)。这些假定中,0 均值假定并不关键,只有在对截距进行估计时才有影响。方差不变的假定有时可以放松,以允许不同时点 t 上的方差有所差异。值得提醒的是,任意时点的 ε_{it}都与任

何其他时点的x_{it}独立，这意味着x_{it}是严格的外生(strictly exogenous)变量。这一假定在某些情况下可以有所松动，但因此涉及的问题绝非无足轻重，也不单纯是技术性问题。在第六章我将讨论其中的一些问题。

至于 α_i，在传统的固定效应分析中，它被假定代表 n 个固定参数，它们要么可以被直接估计出来，要么可以通过某种方式从估计方程中消除掉。如第 1 章已经提到的，在这一章，我们将采用一种新的策略来理解 α_i，假定它代表的是一套随机变量。尽管我们会假定 α_i 和 ε_{it} 在统计上相互独立，但我们允许 α_i 与时变解释向量x_{it}之间的任意相关。另外，如果我们并不关心 γ，也可以允许 α_i 与z_i之间任意相关。此种相关的纳入把固定效应方法与随机效应方法区别开来，同时也才让我们敢说固定效应方法“控制”了不随时间变化而变化的不可观测变量(unobservables)。此刻，我们还不需要对 α_i 的均值和方差作任何假定。

两期数据（固定效应分析）

当变量只被观察两次($T=2$)时，方程 2.1 的估计非常简单。对应的两个方程分别为：

$$y_{i1}=\mu_1+\beta x_{i1}+\gamma z_i+\alpha_i+\varepsilon_{i1}$$

$$y_{i2}=\mu_2+\beta x_{i2}+\gamma z_i+\alpha_i+\varepsilon_{i2} \qquad [2.2]$$

从第二个方程中减去第一个方程，我们就得到了“一阶差分”(first difference)方程：

$$y_{i2}-y_{i1}=(\mu_2-\mu_1)+\beta(x_{i2}-x_{i1})+(\varepsilon_{i2}-\varepsilon_{i1}) \qquad [2.3]$$

这一方程可以被改写成：

$$\Delta y_i=\Delta\mu+\beta\Delta x_i+\Delta\varepsilon_i \qquad [2.4]$$

其中，Δ 表示差分值(difference score)。注意，α_i 和 γz_{it} 被从方程中“差分掉”了。这样，我们就不用再担心 α_i 及其与 Δx_i 之间的可能相关了。当然，从另一方面来讲，我们也就失去了估计 γ 的机会。由于x_{i1}和x_{i2}分别都与 ε_{i1}和 ε_{i2}无关，

可推出 Δx_i 与 $\Delta \varepsilon_i$ 也相互独立。这意味着通过对差分进行一般最小二乘(OLS)回归就能得到 β 的无偏估计。

现在让我们将这一方法用于实际数据。我们的样本来自全美青少年跟踪调查(National Longitudinal Survey of Youth,简称 NLSY; Center for Human Resource Research, 2002)。[1]从原本要大得多的数据集中,我抽取了一个只包含 581 个小孩的子样本,他们在 1990 年、1992 年及 1994 年都接受过调查。第一步,我们只考虑三个在三次调查中都得到测量的变量:

ANTI　反社会行为(取值范围 0 到 6)

SELF　自信水平(取值范围 6 到 24)

POV　如果家庭贫困则编码为 1,否则为 0

此刻,我们先忽略中间一年(1992 年)的观察记录,只使用 1990 年和 1994 年的数据。分析的目标是对以 ANTI 为因变量[2],SELF 和 POV 为自变量的线性方程进行估计:

$$\text{ANTI}_t = \mu_t + \beta_1 \text{SELF}_t + \beta_2 \text{POV}_t + \alpha + \varepsilon_t,\ t = 1,\ 2 \qquad [2.5]$$

通过如此表达这一模型,我们假定了某种特定方向的因果关系,具体而言,是 SELF 和 POV 影响 ANTI,而不是反过来。我们还假定了因果效应是同期发生的(SELF 和 POV 不存在时滞效应)。这两个假定在第 6 章中将会被放宽。最后,我们假定 β_1 和 β_2 在两个时期是一样的,不过这一假定很快就会被放宽。相反,我们让截距 μ_t 在各个时期有所不同,允许非 SELF 或 POV 变化结果的反社会行为的平均水平在不同时期有所变化。

作为开始,我们先用一般最小二乘回归分别对两个时期估计方程 2.5。结果呈现在表 2.1 的头两列。并不意外的是,在两个年份的数据中,贫穷都与较高水平的反社会行为有关,而自信与较低的反社会行为水平相关。两个年份的回归系数都非常相似。

这两个回归都没有对非时变变量(如性别、民族等)进行任何控制。但是,

通过对差分进行回归，而不是将这类变量纳入方程，我们就能控制所有非时变自变量。对于每个小孩及每个变量，我们都用1994年的取值减去1990年的取值，然后用ANTI差分对SELF差分和POV差分做回归。POV是虚拟变量，似乎不宜用一个值减去另一个值。但事实上，在这方面，虚拟变量可像任何其他变量一样处理。

差分回归的结果在表2.1的最后一列。尽管方程是以差分的形式估计的，但回归系数的解释和直接估计2.5式时一样。它们表示某一年的每个变量对当年因变量取值的影响。对自信水平这一变量来说，差分方程估计得到的系数处于两个年份分开估计得到的系数之间，并且仍然高度显著。而贫困状况的回归系数则要小很多，并且在统计上不再显著。

表2.1 反社会行为对自信水平及贫困状况的OLS回归

	1990		1994		差分值	
	系　数	标准误	系　数	标准误	系　数	标准误
截距	2.357**	0.384	2.888**	0.447	0.209**	0.063
SELF	−0.050**	0.019	−0.064**	0.021	−0.056**	0.015
POV	0.595**	0.126	0.547**	0.148	−0.036	0.128
R^2	0.05		0.04		0.02	

注：** $p<1$。

固定效应估计结果与用其他方法估计得到的结果差异巨大的情况是十分常见的。在这个例子里，可能的解释之一是，两个年份分开回归得到的贫困效应估计是虚假的，反映的是贫困与某些影响反社会行为的非时变变量之间的相关。

当然，结论不能下得太过草率。只要常规回归产生的系数显著，而固定效应回归产生的系数不显著，就存在两种可能的解释：(a)固定效应系数从大小上看要小得多，或者(b)固定效应标准误大得多。正如前文已经提过的，固定效应回归系数的标准误通常比其他模型的标准误大，尤其是在自变量的历时变化很小时。事实上，贫困状况的变异主要存在于女孩之间，仅有24%的女孩在1990到1994年期间脱离或陷入贫困境地(即发生过个体内的变化)。

不过，差分方程中贫困状况回归系数的标准误差不多与 1990 年的一样，比 1994 年的甚至还小。因此，变异不足在这里并不成为问题。看来，在控制了非时变变量之后，贫困状况的影响确实有较大的衰退。总的经验是，一旦固定效应方法得到的 P 值与其他方法显著不同时，一定要同时检查系数及其标准误。

最后，可以看到截距 0.209 高度显著。这一系数表示：在自信水平及贫困状况未发生变化的情况下，一个人的反社会行为从时间 1 到时间 2 的变化量估计值。

两期数据差分法的扩展

方程 2.1 对应的基本固定效应模型可以被扩展，以允许 x 和 z 的影响在时间上存在差异。在两期数据情况下，我们可以将上述方程改写成每个时期的系数完全不同的方程：

$$
\begin{aligned}
y_{i1} &= \mu_1 + \beta_1 x_{i1} + \gamma_1 z_i + \alpha_i + \varepsilon_{i1} \\
y_{i2} &= \mu_2 + \beta_2 x_{i2} + \gamma_2 z_i + \alpha_i + \varepsilon_{i2}
\end{aligned} \tag{2.6}
$$

取一阶差分，合并同类项，得到：

$$
\begin{aligned}
y_{i2} - y_{i1} = (\mu_2 - \mu_1) + \beta_2(x_{i2} - x_{i1}) + (\beta_2 - \beta_1)\, x_{i1} \\
+ (\gamma_2 - \gamma_1)\, z_i + (\varepsilon_{i2} - \varepsilon_{i1})
\end{aligned} \tag{2.7}
$$

这一方程可以被改写成：

$$\Delta y_i = \Delta\mu + \beta_2 \Delta x_i + \Delta\beta x_1 + \Delta\gamma z_i + \Delta\varepsilon_i$$

关于这一方程，有三点值得注意。第一，和以往一样，α_i 被差分掉了，因此我们无需担心它的潜在干扰。第二，z 没有被消除，并且其系数向量等于两个时点对应系数向量之差。由此我们知道，回归系数在不同时期发生改变的非时变变量必须被明确地纳入方程。固定效应只能对不随时间变化而变化的非时变变量的影响进行控制。第三，现在，方程含有自变量x_1，并且其回归系数等于两时期对应系数之差。因此，对于 z 和x_1 来说，检验它们的系数等于 0 与检

验 $\beta_1 = \beta_2$ 或 $\gamma_1 = \gamma_2$ 等价。

下面我们尝试将这一模型用于 NLSY 数据。该数据集还包括如下非时变变量，我们将把它们作为可能的解释变量来检查：

BLACK	如果小孩是黑人则等于1，否则为0
HISPANIC	如果小孩是西班牙裔则等于1，否则为0
CHILDAGE	小孩在1990年的年龄
MARRIED	如果其母亲在1990年处于在婚状态则等于1，否则为0
GENDER	女孩等于1，男孩等于0
MOMAGE	小孩出生时母亲的年龄
MOMWORK	如果母亲1990年处于在业状态则为1，否则为0

前两个变量 BLACK 和 HISPANIC 分别代表一个三分变量的两个类别，参照组为非西班牙裔白人。这7个变量将和自信水平及贫困状况的差分一起被纳入到反社会行为的差分方程中。一同被纳入模型的还有1990年测量的自信水平及贫困状况。

表 2.2 扩展差分模型的 OLS 估计

	系 数	标准误	P值
截距	−0.550	1.360	0.6859
SELF 差异分	−0.060	0.020	0.0024
POV 差异分	0.031	0.156	0.8446
1990年的 SELF	−0.018	0.025	0.4826
1990年的 POV	0.121	0.178	0.4991
BLACK	−0.100	0.155	0.5158
SPANIC	0.084	0.164	0.6109
CHILDAGE	0.220	0.107	0.0409
MARRIED	−0.206	0.154	0.1808
GENDER	0.101	0.126	0.4262
MOMAGE	−0.040	0.030	0.1842
MOMWORK	−0.153	0.140	0.2735

呈现在表 2.2 中的结果与表 2.1 中的发现一致。自信水平(差分)的回归系数为－0.06 且高度显著,而贫困状况(差分)的回归系数高度不显著。1990 年的自信水平及 1990 年的贫困状况的系数都不显著,表明自信水平及贫困状况的影响在不同时期保持不变。在 7 个非时变变量中,只有一个——小孩在 1990 年的年龄——统计显著(勉强显著而已)。这并不表示其他 6 个变量不会影响反社会行为。而是说它们的影响在 1990 年和 1994 年实质上是一样的。

每个个体被观察三期及以上的一阶差分方法

当每个个体被观测时点数等于 3 或更多时($T > 2$),如何拓展我们刚刚考虑过的方法,并不那么显而易见。在上述 NLSY 数据中,我们实际上有 3 个年份的数据——1990, 1992 和 1994 年。一种可能的办法是,建立并估计两个一阶差分方程。从等式 2.2 开始,我们有:

$$y_{i2} - y_{i1} = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$
$$y_{i3} - y_{i2} = (\mu_3 - \mu_2) + \beta(x_{i3} - x_{i2}) + (\varepsilon_{i3} - \varepsilon_{i2}) \qquad [2.8]$$

这两个方程可以用 OLS 方法分开进行估计,且都能得到对 β 的无偏估计。表 2.3 前面两大列给出了 NLSY 数据的这一结果。两个差分方程中自信水平的系数都为负,大小几乎一致,并且都高度显著。贫困状况在两个方程中都很不显著。截距表示在控制了上述两个变量后,反社会行为水平从一个时期到下一个时期发生的变化。尽管在两个为期两年的时段中,反社会行为都有所增长,但只有 1992 年到 1994 年的变化在统计上是显著的。

假定 β 系数在不同时期保持不变,则应该对这两个方程进行同时估计,以获得最佳效率。这可以通过创建一个每个人有两条记录的统一数据集来实现,一条记录包含第一个方程所用的差分值,另一条则包括第二个方程所需的差分值。另外还有一个虚拟变量将第一条记录与第二条记录区别开来。而且,还需要有一个变量,该变量中,同一个人的两条记录有着相同的

ID 号。

表 2.3 的第三大列给出了对这一包含 1162 条记录的合并数据集应用 OLS 回归得到的拟合结果。不出所料，自信水平和贫困状况的回归系数取值处于前两列对应系数取值之间。不过，由于使用了更多的信息，(该结果的)标准误要稍微小一些。该模型的截距可以被解释为 $\mu_2-\mu_1$ 的估计值，而方程虚拟变量的系数则是对 $(\mu_3-\mu_2)-(\mu_2-\mu_1)$ 的估计。两者都是正值，这说明反社会行为从时间 1 到时间 2 有所增加，并且从时间 2 到时间 3 期间增加得更快。不过，两者在统计上都不显著。

表 2.3　用自信水平及贫困状况解释反社会行为的一阶差分回归

	1992—1994 OLS		1990—1992 OLS		组合 OLS		组合 GLS	
	系　数	标准误	系　数	标准误	系　数	标准误	系　数	标准误
截距	0.71**	0.059	0.040	0.053	0.045	0.056	0.05	0.056
SELF	−0.072**	0.016	−0.039**	0.014	−0.055**	0.010	−0.055**	0.010
POV	0.216	0.136	0.197	0.133	0.213	0.095	0.139	0.094
方程虚拟变量					0.122	0.080	0.122	0.094

注：** $p<0.01$

尽管组合 OLS 回归估计是无偏的，但它忽视了 $\varepsilon_2-\varepsilon_1$ 同 $\varepsilon_3-\varepsilon_2$ 之间很有可能存在的负向相关，因为它们包含一个共同因素 ε_2，符号却相反。这意味着系数估计的效率可能并不充分，且标准误的估计可能有偏差。这一问题可以通过先估计出误差项之间的相关，然后用广义最小二乘法(generalized least squares，GLS)结合相关办法来解决。

多数综合性统计软件都带有进行 GLS 分析的程序。这类程序通常需要指定一个 ID 变量，以识别哪些记录来自同一个体。这里我用的是 Stata 软件中的 xtreg 命令及 pa 选项，它能以 GLS 方式估计线性模型。[3](本例题的)GLS 估计结果在表 2.3 的最后一列，它们和前一列的 OLS 系数估计值及标准误非常相似。

一阶差分方法很容易就能扩展到每个个体被观察三期以上的情况。如

果每个个体被观察了 T 个时期，就创建 $T-1$ 条记录，每条记录都含有各变量相邻两个时期的差分值。此外，必须有一个变量赋予来自同一个体的所有记录相同的 ID 号，而另一个或一组虚拟变量则将同一个体的不同记录区别开来。然后对全部记录进行回归估计，并利用 GLS 对误差项之间的相关进行修正。除非 T 特别大，如大于 10，否则最好允许误差项相关矩阵处于非结构状态(unstructured)。也就是说，该矩阵允许每对误差项之间的相关都不同。在 T 较大的情况下，更好的做法是(给该矩阵)强加一个简化结构，以减少需要估计的不同相关的数量。更多细节请参见格林的著作(Greene，2000)。

每个个体被观察两期及以上的虚拟变量法

尽管多阶差分法(multiple-difference-score)是估计多期数据固定效应模型的合理方式之一，不过“固定效应”这一称呼通常被保留给另外一种不同的方法，这种方法既可以通过虚拟变量实现，也可以通过创建离均差(mean deviation)的方式来实现。由固定效应法产生的结果与差分法产生的结果虽然经常极其相似，但并不完全相同。在两期数据情况下，两种方法给出的结果完全相同。

虚拟变量法所要求的数据集具有非常不同的结构：(在这种数据集中)每个个体每一时期都有一条记录。以 NLSY 数据为例，要求的数据集中，581 个小孩每个人都有 3 条记录，总共有 1743 条记录。每条记录中同一个时变变量的变量名相同，但取值不同。而所有非时变变量的取值，在同一个体的不同记录上只是简单复制而已。数据集中有一个 ID 变量，来自同一个体的所有记录(在该变量上)取值相同。最后，有一个变量将每个个体的不同时期区分开来。例如在 NLSY 数据中，TIME 变量的取值 1、2、3 对应着 1990、1992、1994。表 2.4 呈现了该数据集的前 15 条记录，对应着开始的 5 个人。

表 2.4 每人 3 条记录的数据集(前 5 个人)

ID	TIME	ANTI	SELF	POV	GENDER
1	1	1	21	1	1
1	2	1	24	1	1
1	3	1	23	1	1
2	1	0	20	0	1
2	2	0	24	0	1
2	3	0	24	0	1
3	1	5	21	0	0
3	2	5	24	0	0
3	3	5	24	0	0
4	1	2	23	0	0
4	2	3	21	0	0
4	3	1	21	0	0
5	1	1	22	0	1
5	2	0	23	0	1
5	3	0	24	0	1

为了使用这种方法，首先必须建立一套虚拟变量，以将数据集中的每个个体区别开来。在我们的例子里，这意味着(需要建立)580 个虚拟变量，以代表 581 个儿童。只要将 ID 变量设置为分类变量，很多统计软件就能自动完成这一任务。如果 TIME 变量也被设置成为分类变量，那就会产生两个虚拟变量，以区别这三个(调查)年份。然后我们就可以用 OLS(一般最小二乘法)来对系数进行估计。事实上，由 ID 变量创建而来的虚拟变量的回归系数是对方程 2.1 中的 α_i 的估计，只是其中之一已被限定等于 0。

我在 Stata 中用 reg 命令完成这一回归[4]，结果在表 2.5 左边一栏。这里只给出了前 9 个虚拟变量的系数。

表 2.5 用自信水平及贫困状况对反社会行为做回归——虚拟变量法

	固定效应			常规 OLS		
	系　数	标准误	p	系　数	标准误	p
SELF	−0.055	0.010	0.00	−0.067	0.011	0.00
POV	0.112	0.093	0.23	0.518	0.079	0.00
TIME_2	0.044	0.059	0.45	0.051	0.090	0.58
TIME_3	0.211	0.059	0.00	0.223	0.091	0.01
ID_2	−0.887	0.819	0.28			
ID_3	4.131	0.811	0.00			
ID_4	1.057	0.819	0.20			
ID_5	−0.536	0.819	0.51			
ID_6	0.040	0.820	0.96			
ID_7	2.170	0.821	0.01			
ID_8	0.910	0.820	0.27			
ID_9	−0.276	0.819	0.74			

将表 2.5 中的结果与表 2.3 最后一列(通过一阶差分法得到的结果)进行比较,可以看到,自信水平的回归系数及标准误看起来几乎一样。虚拟变量法得到的贫困状况的回归系数略微小些,但在两种方法中都很不显著。TIME_2 和 TIME_3 的系数表示与参照组(TIME_1)的对比情况。可以发现,就平均水平而言,反社会行为随着时间的变化有所增长,TIME_3 显著地高于 TIME_1。

为了进行比较,表 2.5 右边部分给出了未放入 580 个虚拟变量时系数的 OLS 估计情况。与两期观察情况下看到的一样,两种方法所得结果的较大差异主要在于常规 OLS 回归中 POV 的系数要大得多,并且高度显著。因此,当我们控制了个体间的所有差异,只关注个体内变化时,贫困对反社会行为的突出影响也就消失了①。另外,标准误的比较也有点意思。POV 系数的标准误在固定效应估计中相对较大,这是没有使用个体间差异信息导致的典型结果。另一方面,对 SELF 及两个 TIME 虚拟变量来说,固定效应标准误比常规 OLS 的要小。为什么会有这种差异呢? 这涉及个体内变异与个体间变异的相对大小问题。对 POV 变量而言,70%的变异在个体之间,而在 SELF 变量中,这一比例只有

① 此处原文为"the apparent effect of poverty on self-esteem",根据上下文应该是作者笔误的结果。因为模型中是用贫困状况解释反社会行为,自信水平并不是因变量。——译者注

53%。[5]至于TIME虚拟变量,所有的变动都是个体内的,个体之间没有差别。(事实上)最适合使用固定效应分析的情况是:时变预测变量的所有变异都是个体内的,而在反应变量上,不同个体之间仍然存在大量差异。

虚拟变量法的问题在于:对全部虚拟变量的系数进行估计,会使计算负担非常重,尤其是在样本量很大的情况下,甚至会超过软件或机器内存的能力范围。幸好,还有另外一种备选的算法——离均差法——能够产生完全一样的结果。唯一不足的是,后者不提供代表不同个体的虚拟变量的系数的估计,不过它们倒很少是我们所关心的。

离均差算法是这样进行的。对于每个个体及每个时变变量(包括反应变量和解释变量),我们都计算该个体在不同时点上的取值的均值:

$$\bar{y}_i = \frac{1}{n_i}\sum_t y_{it}$$

$$\bar{x}_i = \frac{1}{n_i}\sum_t x_{it}$$

其中,n_i 是个体 i 被测量的次数。然后用每个变量的观察值减去每个人对应变量的平均值:

$$y_{it}^* = y_{it} - \bar{y}_i$$

$$x_{it}^* = x_{it} - \bar{x}_i$$

最后,将 y^* 对 x^* 及表示时间效应的变量做回归。这种方法有时被称为"条件"法,因为它把固定效应虚拟变量的系数通过"条件"而去掉了。

如果手工创建这些离均差值,然后利用普通回归程序估计这些系数,你会得到所有这些系数的正确OLS估计。但标准误和p值是不正确的。这是因为常规回归中自由度的计算仅基于设定模型中的自变量数,而实际上它应该包括模型中潜在地使用了代表样本中不同个体的虚拟变量的数目(在NLSY数据中为580)。我们当然可以找到正确计算标准误及p值的公式(Judge, Hill, Griffith & Lee, 1985),但让软件帮你计算岂不更省事。例如,Stata中的xtreg[6]命令就能为固定效应模型提供正确的计算;SAS程序PROC GLM过程步中的ABSORB语句也能做到。

利用 xtreg 命令，我设置了一个固定效应模型(FE 选项)，将 ID 作为识别同一个体记录的变量。结果与表 2.5 前 5 行一样。xtreg 还专门为固定效应模型提供了另外几个统计量：

(1) F 检验，检验所有固定效应虚拟变量系数都等于 0 这一虚无假设是否成立。在这个例子里，p 值比 0.0001 还要小，因此，可以很有把握地拒绝虚无假设。这相当于说，有证据表明存在于个体层面的未被观测的异质性。也就是说，个体之间在反社会行为上存在着稳定的差异，已被观测到的解释变量并不能完全解释这些差异。

(2) 因变量方差中被固定效应(各个 α_i)所解释的比例的估计，这一统计量标着"rho(fraction of variance due to u_i)"。在这个例子里，它的估计值是 0.64。

(3) 固定效应 α_i 与估计的时变自变量线性组合 $\hat{\beta}x_{it}$ 之间的相关系数的估计。在随机效应模型中，这一相关系数被假定等于 0。对于当下这个数据，该相关系数等于 0.068。

(4) 三个确定系数 R^2：个体内确定系数、个体间确定系数及总的确定系数。个体内确定系数就是用离均差变量进行回归时得到的常规确定系数。这里为 0.033。个体间确定系数是各个体的 y 均值与各个体的 y 均值预测值之间的相关系数的平方，在此是 0.041。最后，总的确定系数(0.036)是 y 值本身与 y 值预测值之间的相关系数的平方。这三个确定系数都是用预测值计算得到的，后者基于估计的回归系数，但没有使用固定效应虚拟变量的回归系数。如果将这些系数(在使用虚拟变量法时)纳入进来，这一数据的确定系数将提高到 0.73。

前面已经提过，这种方法的特征是它不能对非时变自变量的系数进行估计。这是很显然的，因为用每个个体的非时变自变量取值(在所有时期都一样)减去其对应的个体内均值后，所有个体取值都将等于 0。应该记住的是，我们仍然控制了所有非时变自变量，尽管它们已被从方程中排除掉了。在下一节，我们将进一步讨论如何检验这些变量的影响本身是否也是非时变的。

在固定效应法中设置与时间的交互作用

在两期数据情况下，我们已经知道如何拓展差分法以让自变量的回归系数在不同时期有所不同。对于时变变量，将时间1时的测量值纳入原差分模型即可。对于非时变变量，则直接纳入模型。在虚拟变量法（或等效的离均差法）中，这一扩展是通过加入各变量与时间的交互项来实现的。

对于有三期调查的NLSY数据，表2.6给出了纳入TIME（当作分类变量处理）与时变变量及非时变自变量的交互作用后的模型结果。由于TIME有3个类别，因此它与每个自变量有两个交互项。注意，模型中不包括非时变自变量的主效应。即使我们试图将它们纳入模型，软件也会都丢掉，因为它们在个体内不存在变异。

表2.6　与时间的交互作用

	系　数	标准误	t	p
TIME_2	0.291	1.245	0.23	0.82
TIME_3	−0.444	1.258	−0.35	0.72
SELF	−0.034	0.016	−2.08	0.04
POV	0.097	0.130	0.75	0.46
TIME_2* SELF	−0.026	0.020	−1.28	0.20
TIME_3* SELF	−0.023	0.021	−1.09	0.28
TIME_2* POV	−0.112	0.152	−0.74	0.46
TIME_3* POV	0.099	0.155	0.64	0.52
TIME_2* BLACK	0.250	0.144	1.74	0.08
TIME_3* BLACK	−0.110	0.144	−0.77	0.44
TIME_2* HISPANIC	0.190	0.154	1.23	0.22
TIME_3* HISPANIC	0.075	0.153	0.49	0.62
TIME_2* CHILDAGE	0.076	0.100	0.76	0.45
TIME_3* CHILDAGE	0.227	0.100	2.26	0.02
TIME_2* MARRIED	−0.095	0.143	−0.67	0.51
TIME_3* MARRIED	−0.176	0.143	−1.23	0.22
TIME_2* GENDER	0.041	0.118	0.35	0.73
TIME_3* GENDER	0.107	0.118	0.91	0.37
TIME_2* MAMAGE	−0.027	0.028	−0.96	0.34
TIME_3* MOMAGE	−0.042	0.028	−1.52	0.13
TIME_2* MOMWORK	0.0137	0.131	1.05	0.29
TIME_3* MOMWORK	−0.144	0.130	−1.11	0.27

对于每一个交互项，t 统计量检验的是系数在 Time2 或 Time3 时是否与在 Time1 时不同。在 18 个交互项中，只有一个(TIME_3* CHILDAGE)统计显著($p = 0.024$)。对于该交互项，系数 0.227 表明 Time3 时 CHILDAGE 的系数比 Time1 时要高 0.227。当然，在检验多达 18 个的情况下，即使没有任何实质性依据，我们也可以大胆地赌一把，认为其中至少会有一个显著。不过，在检验 18 个交互项都等于 0 的同时检验(simultaneous test)中，p 值等于 0.15。[①]

与随机效应模型的比较

固定效应模型的一个非常流行的替代者是随机效应或混合模型(mixed model)。这一模型是在我们用于固定效应模型的同一方程上发展出来的：

$$y_{it} = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i + \varepsilon_{it} \qquad [2.9]$$

最关键的区别在于，现在我们不把 α_i 当作一套固定数字(等价于把 α_i 看作随机的，但与x_{it}之间存在所有可能的相关)，而假定 α_i 是一套有着特定概率分布的随机变量。例如，通常假定每个 α_i 都服从 0 均值且等方差的分布，并且与方程右边的其他所有变量都保持独立。

现在有很多软件可以用来估计随机效应模型。SAS 可以通过 MIXED 程序进行。Stata 中的 xtreg 命令在默认情况就能进行随机效应模型的 GLS 估计。表 2.7 给出了 xtreg 命令产生的包含非时变变量及不含非时变变量的随机效应模型的估计结果。

能够纳入非时变变量是随机效应模型与固定效应模型最显著的差别。不过，这里我们发现纳入此种变量并不会使时变预测变量的系数发生太大改变，不管是自信水平还是贫困状况。

与表 2.5 中的常规 OLS 回归估计相同而与固定效应估计不同的是，随机效应估计中这两个变量的系数都是高度显著的。随机效应模型与常规 OLS 方法

---

① 这说明整体而言，18 个交互项并没有显著的解释效果，纳入交互项后模型没有显著的改善。——译者注

相似，并不足为怪。如果 α 与其他所有变量都不相关这一随机效应假定正确的话，这两种方法都能产生对方程 2.9 的系数的一致估计（因而也是近似无偏的估计）。但如果这一假定并不正确，那这两种方法的估计都将有偏。

表 2.7　随机效应模型的 GLS 估计

	系　数	标准误	p	系　数	标准误	p
SELF	−0.062	0.009	0.00	−0.060	0.009	0.00
POV	0.247	0.080	0.00	0.296	0.077	0.00
TIME_2	0.047	0.059	0.42	0.047	0.059	0.42
TIME_3	0.216	0.059	0.00	0.216	0.059	0.00
BLACK	0.227	0.126	0.07			
HISPANIC	−0.218	0.138	0.11			
CHILDAGE	0.088	0.091	0.33			
MARRIED	−0.049	0.126	0.70			
GENDER	−0.483	0.106	0.00			
MOMAGE	−0.022	0.025	0.39			
MOMWORK	0.261	0.115	0.02			

为何 POV 变量在随机效应模型中高度显著而在固定效应模型中非常不显著呢？如早前已经解释过的一样，一旦一个系数在随机效应模型中显著而在固定效应模型中不显著，首先要做的就是比较两者的标准误。因为固定效应模型的标准误通常要比随机效应模型的标准误大得多，仅此一点，就能解释它的 p 值为什么会较大。当然，在这里，固定效应模型中 POV 变量的标准误虽然稍大（0.09 比 0.08），但并不足以解释上述显著水平上的差异。即使我们将 0.09 替换为 0.08，固定效应模型中 POV 的系数仍然不会显著。显然，主要差异在于两个系数的大小不同，在固定效应模型中是 0.11，在随机效应模型中是 0.25（或 0.30，取决于非时变变量是否得到控制）。对这一差异最可能的解释是，存在着某些不可观测的变量能够“解释”我们（在随机效应模型中）看到的贫困状况与反社会行为之间的相关。一旦这些不可观测的变量通过固定效应模型被控制了，上述两个变量之间的相关也就消失了。

这里的关键在于，与一般流行的观念不同，估计一个随机效应模型并不能真正"控制"未被观测的异质性。这是因为，常规随机效应模型假定观测变量与未被观测的变量之间不存在相关。相反，固定效应模型允许非时变自变量与时变自变量之间的任何相关。不过，这样做的代价是，当这些相关确实为零时，固定效应模型将失去一些效率。

已有的研究显示，随机效应模型实际上只是固定效应模型的一个特例(Mundlak，1978)。也就是说，如果以方程 2.9 的常规随机效应模型作为开始，然后允许 α_i 与 x_{it} 变量之间的所有可能相关，你将得到固定效应模型的等价物。通常而言，一旦要在两个相互嵌套的模型(其中一个是另一个加上某些限制条件的结果)之间做出选择，都会存在偏差与效率之间的得失权衡问题。较简单的模型(随机效应模型)可以得到更有效率的估计，但如果加在模型上的限制条件是错误的，那这些估计就可能是有偏的。较为复杂的模型(固定效应模型)不那么容易产生偏差，但代价是抽样变异性(sampling variability)会相对较大。

在此得失权衡面前，如果有一种统计检验能对随机效应模型与固定效应模型进行比较，那将大有用处。此种检验能够帮助我们判定随机效应方法所带来的偏差是小到足以忽略的程度，还是大到我们不得不选择限制条件更少的固定效应模型。这些检验中，最有名的是 Hausman 检验(1978)，该检验的虚无假设是随机效应系数与固定效应系数相同。[7]这种检验在很多统计软件中都能找到。对于手头这个例子，最直接的检验是对表 2.5 中的固定效应系数与表 2.7 左边部分的随机效应系数进行比较，后者控制了几个非时变变量。我在 Stata 中用 xtreg 命令的这一检验得到的 p 值等于 0.04，这一证据并不支持随机效应模型，而比较倾向于固定效应模型。在下一节，我将介绍另外一种效果比目前 Stata 中采用的 Hausman 检验更好的检验方法。

混合（模型）法（A Hybrid Method）

现在来考虑如何将固定效应模型和随机效应模型的某些优点综合起来。前面我们已经看到，估计固定效应模型的方法之一是将所有变量都表达成与个

体均值的离差，然后对这些离均差变量运用 OLS 回归。在混合法中，时变 x 变量再次被处理成为与个体均值的离差，但是反应变量 y 没有。而且与前面的固定效应方法不同，现在我们将非时变变量 Z 也纳入到模型里。此外，我们还将表示每个时变变量个体内均值的变量(同样是非时变变量)也纳入模型。最后，我们不采用 OLS 回归，而是估计一个随机效应模型，以保证标准误能够反映同一个体的多个观测记录之间的相依性。[8]

表 2.8 给出了针对 NLSY 数据的分析结果。DSELF 和 DPOV 是离均差变量。MSELF 和 MPOV 是个体内均值。首先应该注意的是，表中 DSELF，DPOV 以及两个时间虚拟变量的回归系数和标准误与我们在表 2.5 的固定效应方法中看到的完全一样。因此，我们又有了一种产生固定效应估计的方法。[9]实际上，对于 DSELF 和 DPOV 来说，不管方程中放进什么非时变变量，即使我们将 MSELF 和 MPOV 以及其他非时变变量都删除，时变变量的系数和标准误都将保持不变。当然，从这一混合方法中我们得到的是对非时变变量效果的估计，而这是通过常规的固定效应方法所不能得到的。

表 2.8　混合法估计结果

	系　数	标准误	p
DSELF	−0.055	0.010	0.00
DPOV	0.112	0.093	0.22
MSELF	−0.090	0.022	0.00
MPOV	0.616	0.157	0.00
BLACK	0.111	0.132	0.40
HISPANIC	−0.280	0.139	0.04
CHILDAGE	0.086	0.091	0.35
MARRIED	−0.128	0.128	0.32
GENDER	−0.508	0.107	0.00
MOMAGE	−0.011	0.025	0.65
MOMWORK	0.164	0.119	0.17
TIME_2	0.044	0.059	0.45
TIME_3	0.211	0.059	0.00

在多层模型文献中(Bryk & Raudenbusch，1992；Glodstein，1987；Kreft & De Leeuw，1995)，将各时变变量减去个体内均值的做法叫作按组均值对中(group mean centering)。尽管大家都知道对中后将产生非常不同的结果，但这类文献仍没有将其与固定效应模型联系起来，也没有认识到按组均值对中将控制所有非时变预测变量。

均值变量 MSELF 和 MPOV 的估计系数本身并不特别具有启发性。但将这些变量纳入模型很重要，原因有二：第一，它们可以帮助我们得到更好的有关其他非时变变量效果的估计。将 MSELF 和 MPOV 排除，意味着我们没有完全控制这些变量。第二，将它们的系数与离均差变量 DSELF 和 DPOV 的系数进行比较，能够让我们对事情有更深的了解。如果随机效应模型的假定是正确的(即 α_i 项与 x 变量无关)，那么每个变量对应的离均差变量与均值变量的系数应该一样(除去抽样变异后)。对于 DSELF 与 MSELF 来说，这确实相差不远。但 MPOV 的系数比 DPOV 的系数要大得多。实际上，在我们估计常规随机效应模型时，得到的 SELF 和 POV 的系数就是这些"(个体)内"系数和"(个体)间"系数的加权平均数。这进一步意味着我们可以通过检验这两对系数之间的相等性来检验随机效应模型与固定效应模型的差异(这就是可以替代前面讨论过的 Hausman 检验的另一检验)。这在 Stata 中做起来非常容易(使用的是 Wald 检验)，所得 p 值为 0.007，可算是反对随机效应模型的鲜明证据(clear-cut evidence)。

混合法的另一诱人之处，在于它可以实现在常规的固定效应估计方法中不易实现的多种有趣的拓展。到目前为止，我们讨论的随机效应模型都还只是随机截距模型。我们还有可能估计随机斜率模型。例如，我们不再假定 DSELF 的系数对于每个人都一样，而假定其为一个随机变量，然后估计其均值与标准差。此种模型通过 SAS 中的 MIXED 子程序或者 Stata 中的 xtmixed 命令很容易就能搞定。后一命令产生的 DSELF 系数的平均值的估计值为－0.005。该系数对应的标准差估计值为 0.070，是其标准误 0.024 的两倍以上，这强烈地表明：DSELF 的影响在不同个体上存在差异。

通过使用混合法，还可以估计含有更复杂的误差结构的模型，例如，三层结

构或自回归结构(three-level structure or autoregressive structure),而不仅仅是常规固定效应模型所隐含的简单结构。关于此类模型的更多信息,可以参看辛格和威利特的著作(Singer and Willett, 2003)。

总结

可以看到,共有几种等价的方法用来估计定量反应变量线性固定效应模型:

(1) 如果每个个体只被观察了两期,先对所有时变预测变量创建差分值。然后,对反应变量的差分值对预测变量的差分值做 OLS 回归。

(2) 不管观察了多少个时期,转换数据结构以使每个个体在每一观察期都有一条记录。然后进行 OLS 回归,回归时纳入代表每个个体的虚拟变量组(缺省其中的一个)。

(3) 对于方法 2 中的数据结构,将所有时变变量转变为相对于个体内均值的离差值。然后对这些离差值做 OLS 回归(并修正标准误统计检验及 p 值)。通过 Stata 中的 xtreg 命令可以很方便地完成。

(4) 对于方法 2 中的数据结构,只将自变量转变为相对于个体内均值的离差值。然后估计一个随机效应模型,模型中的自变量中同时包括(各时变变量的个体内)均值以及相对这些均值的离差值。

在这些方法中,第 4 个最为灵活。它具有其他一个或多个方法所不具备的如下能力:

纳入在个体内不存在变异的预测变量;

将固定效应与随机效应假定进行检验;

为存在个体内变异的自变量提供随机系数估计;

容许更为宽松的误差结构。

无论使用何种运算方法，固定效应模型都能高效地控制所有非时变预测变量，不管是得到测量的还是没有被测量的。这是其与随机效应模型相比的主要吸引力所在。不过，固定效应模型的一个重要假定在于，非时变预测变量在各个时期必须有着相同的影响。那些作用效果在各个时期并不恒定的变量必须被明确地纳入模型（才能得到控制）。另外，当然，固定效应方法对未被测量的时变变量没有任何控制。

第 3 章 | 固定效应 Logistic 回归

在这一章，我们将了解到如何将上一章的固定效应方法一般化，以适用于分类反应变量。为了探索这些方法，我们将用一个同样来自全美青少年长期跟踪调查(NLSY)的数据集。这一数据集有 1151 名自 1979 年开始每年都被访问一次、连续访问了 5 年的青少年女孩。反应变量 POV1-POV5 是二分变量：在这 5 年期间的每一年，判断女孩所在家庭根据美国联邦的标准是否处于贫困状态，贫困则编码为 1，非贫困则编码为 0。我们的自变量如下：

AGE	第一次被访时的年龄
BLACK	如果受访者为黑人则编码为 1，否则为 0
MOTHER	如果受访者目前至少有一个孩子则编码为 1，否则为 0
SPOUSE	如果受访者目前与配偶生活在一起则编码为 1，否则为 0
SCHOOL	如果受访者目前为在校注册学生则编码为 1，否则为 0
HOURS	调查所在周已工作的小时数

开始两个变量是非时变变量，而后面四个变量在每次调查时都可以不同。

现在我们处理的已经不是线性模型了，而是 Logistic 回归模型，与方程 2.1 相似，我们的基础模型是：

$$\log\left(\frac{p_{it}}{1-p_{it}}\right)=\mu_t+\beta x_{it}+\gamma z_i+\alpha_i,\ t=1,\ 2,\ \cdots T \qquad [3.1]$$

其中，p_{it}是响应变量等于 1 的概率。如以前一样，x_{it}是时变预测变量向量，z_i 是非时变预测变量向量，α_i 表示所有未被观测的恒定变量的综合影响。在这一

章，我们将把 α_i 看作是一套固定的常量，每个个体都有一个。不过，这相当于假定 α_i 是随机的且对 α_i 与 x_{it} 之间的相关不做任何限制。

两期数据（固定效应分析）

在第 2 章中，我们看到两期数据情况下，固定效应线性模型可以通过计算所有变量的差分值，然后运用一般最小二乘回归的方式进行估计。对于 Logistic 回归，类似的做法同样是可能的，但有一些重要的区别。

表 3.1 第 1 年与第 5 年贫困状况的交互分类情况

第 1 年的贫困状况	第 5 年的贫困状况		
	0	1	合 计
0	516	234	750
1	211	190	401
合 计	727	424	1 151

下面我们对 NLSY 数据做一个固定效应 Logistic 回归，先忽略第 2、3、4 年，只关注年份 1 和年份 5。尽管这两个年份间贫困的边缘分布改变得很小，但仍然有 234 个女孩的家庭陷入贫困状态，同时有 211 个女孩脱离了贫困。

为了进行固定效应 Logistic 回归，我们先将 706 个（贫困状况）在 5 年内未发生变化的女孩排除在外。这是因为固定效应模型只使用个体内变异，对于这些女孩而言，其响应变量并不存在个体内的变化。因此，我们只剩下 445 个贫困状况发生过改变的女孩。在这一缩减后的样本上，我们令 p_i 表示 POV5 = 1 的概率，等于一个女孩从状态 0 转变成状态 1，而不是从 1 转变为 0 的概率。接下来我们再用常规的最大似然法估计这一模型：

$$\log\left(\frac{p_i}{1-p_i}\right)=(\mu_2-\mu_1)+\beta(x_{i2}-x_{i1}) \qquad [3.2]$$

也就是说，我们将 POV5 作为因变量，将时变预测变量的差分值作为自变量，作 Logistic 回归。根据下一节的解释，这实际上是条件最大似然估计的一种

形式。与线性模型中一样，z_i 和 α_i 都被从方程中消除。[10]

表 3.2 呈现了 3 个回归模型的估计结果，是用 Stata 中的 logit 命令估计得到的。模型 1 只包括时变预测变量的差分值。可以看到成为母亲将增加陷入贫困的风险，而与配偶居住及工作的小时数越多将降低风险。请再次记住，这些估计都已经控制了所有非时变变量。将是否为母亲这一变量的系数(0.730)取指数，我们得到 2.08。这告诉我们：一个女孩一旦有了第一个孩子，其陷入贫困的风险将会翻倍。截距 0.539 可以被解释为一个在所有时变自变量上都未发生变动的女孩的贫困对数发生比从年份 1 到年份 5 的变化。取指数后，我们得到发生比为 1.71——也就是说，从年份 1 到年份 5，提高了 71%。

表 3.2 两期数据的差分值 Logistic 回归

	模型 1		模型 2		模型 3	
	系 数	标准误	系 数	标准误	系 数	标准误
截距	0.539**	0.162	4.899**	1.644	3.052	1.826
DMOTHER	0.730**	0.250	0.744**	0.254	0.909**	0.270
DSPOUSE	−1.002**	0.283	−1.032**	0.292	−1.022**	0.301
DSCHOOL	0.343	0.212	0.339	0.218	0.639*	0.251
DHOURS	−0.0339**	0.0061	−0.0339**	0.0062	−0.0339**	0.0068
BLACK			−0.526*	0.216	−0.662**	0.226
AGE			−0.258*	0.103	−0.196	0.111
MOTHER1					0.457	0.460
SPOUSE1					0.442	0.726
SCHOOL1					1.184	0.471
HOURS1					−0.0024	0.0128

注：* $0.01 < p < 0.5$ ** $p < 0.01$。

模型 2 加入了两个非时变变量 BLACK 和 AGE，两者都有着显著的负向作用。这些变量的系数可以被解释为与时间的交互作用。因此对这两个变量来说，它们对于陷入贫困的风险的影响(从量上说)在年份 5 时要比在年份 1 时小。或者，这些系数也可以解释为历时变化的速率在不同子群体中是如何的不同。更具体的，对于一个从年份 1 到年份 5 期间时变自变量都未发生变化的女孩来说，这 5 年期间陷入贫困的对数发生比的变化可以表达为：

$$4.899 - 0.526 \times \text{BLACK} - 0.258 \times \text{AGE}^{①}$$

因此,对于一个 14 岁的在各项其他自变量上都未发生变化的非黑人女孩来说,其陷入贫困的对数发生比的预测改变量为 1.29。等价地说,其陷入贫困的发生比将变为原来的 exp(1.29) = 3.63 倍。而黑人或在年份 1 时年龄更大的女孩贫困的增长率要更低。

模型 3 增加了在年份 1 时测量的时变变量。如在第二章一样,这些变量的回归系数可以被解释为每个变量从年份 1 到年份 5 时作用的变化量,也就是说,与时间的交互作用。这些变量中只有一个,SCHOOL1,统计显著。这样,我们可以说,从年份 1 到年份 5 期间,是否为在校学生对女孩陷入贫困的对数发生比的影响增加了 1.184。而 DSCHOOL 的系数 0.639 是在校学生身份在年份 5 时的估计作用。由此看来,在校学生身份在年份 1 时为负作用(0.639 − 1.184),而在年份 5 时为正作用。

总的来说,两期数据 Logistic 模型与第二章中两期数据的线性回归是非常相似的。最大的不同在于:Logistic 方法要求将因变量上未发生变动的个体排除在样本之外。对于因变量,我这里使用的是时期 2 的反应变量,这看起来与线性固定效应模型时有所不同。但是如果我不这么用,而用时期 2 的取值减去时期 1 的取值,那我得到的因变量取值将会是 1 和 −1 而不是 1 和 0,但这二者实际上是同一个东西。[11]

三期及多期数据(固定效应分析)

如何才能将这种方法扩展到可以使用全部 5 年而不仅仅是第 1 年和第 5 年数据提供的信息呢?在第 2 章,我们是通过如下方式来实现的:为每个个体在每一个被测量的时期创建一条单独记录,将这些记录合成一个数据集,然后估计一个含有与各个体对应的虚拟变量的线性回归。另外一种方法同样是使用单

① 原书中为 4.899−0.526×BLACK−2.58×AGE,但是根据表 3.2 的输出结果,AGE 的系数应该为 − 0.258。—— 译者注

一个体对应多条记录的数据形式，但避免使用虚拟变量，而是将各个变量表达为相对于个体内均值的离差的形式。尽管这两种方法会产生同样的结果，但第一种方法确切地讲是无条件最大似然法，而后者为条件最大似然法。

条件最大似然法和无条件最大似然法对二分结果变量的 Logistic 回归都是可用的，但在这里两者并不产生相同的结果。与在线性情况下一样，无条件最大似然法同样是通过为每个个体创建多条记录，然后估计一个含有标识各个体的虚拟变量的常规 Logistic 回归来实现的。不幸的是，这种方法产生的系数估计是有偏的(Hsiao, 1986)。事实上，在两期观察情况下，系数估计值恰好是其本来取值的两倍(Abrevaya, 1997; Hsiao, 1986)。导致此种偏差的原因就是所谓的伴随性参数问题(*incidental parameters problem*)(Kalbfleisch & Sprott, 1970; Lancaster, 2000)。也就是说，样本规模一旦增加，参数(尤其是与各个体对应的虚拟变量的系数)的数量也会直接增加，从而违背了最大似然估计的渐近理论的重要前提之一。

解决方案就是采用条件似然法，这一方法使 α_i 参数被"条件出(conditions out)"似然方程①(Chamberlain, 1980)。它是通过将似然方程限定在每个个体被观察的事件总数一定的条件上来实现的。从效果上讲，每个人对似然函数的贡献，就是对如下问题的回答：如果一个女孩在 5 年内有两年处于贫困状态，那么这一事件发生在，比方说年份 2 和年份 4(当实际发生时)，而非另外年份 9 对可能的年份组合之一的概率是多大？这些条件概率不包括 α_i 参数。此种条件似然法只适用于二分类反应变量的 Logistic 回归，不能用于其他"连接"函数，如 Probit 或互补双对数函数。

很多统计软件都能对 Logistic 回归的此种条件的似然值进行最大化估计。在 Stata 中，通过 xtlogit 或者 clogit 命令都可以实现。这些程序所要求的数据形式与第 2 章中讨论多期数据与 xtreg 命令时描述的一样。每一个体在每个被观察的时期都有一条记录，来自同一个体的所有记录都有同样的 ID。非时变变量的取值在同一个体的不同记录中是一样的。如果将此种方法应用于两期数据，其产生的结果将与上面描述的差分法相同。对于这个 5 年份示

---

① 实际上就是在方程中被抵消掉。——译者注

例数据，表 3. 3 给出了工作数据集中的前 15 条观察记录。这些记录来自 3 个女孩，每个人被观察了 5 年。

表 3.3　少女贫困问题数据集的前 15 个观察记录

观测值	ID	YEAR	POV	MOTHER	SPOUSE	SCHOOL	HOURS	BLACK	AGE
1	22	1	1	0	0	1	21	0	16.00
2	22	2	0	0	0	1	15	0	16.00
3	22	3	0	0	0	1	3	0	16.00
4	22	4	0	0	0	1	0	0	16.00
5	22	5	0	0	0	1	0	0	16.00
6	75	1	0	0	0	1	8	0	17.00
7	75	2	0	0	0	1	0	0	17.00
8	75	3	0	0	0	1	0	0	17.00
9	75	4	0	0	0	1	4	0	17.00
10	75	5	1	0	0	1	0	0	17.00
11	92	1	0	0	0	1	30	0	16.00
12	92	2	0	0	0	1	27	0	16.00
13	92	3	0	0	0	1	24	0	16.00
14	92	4	1	1	0	0	31	0	16.00
15	92	5	1	1	0	0	0	0	16.00

Stata 的 xtlogit 命令对固定样本跟踪数据拟合 Logistic 回归可以采用三种不同的方法：固定效应法（条件似然），随机效应法及广义估计方程法（generalized estimating equation）。表 3. 4 给出了对该少女贫困数据应用全部三种方法得到的结果。前面两列呈现的是固定效应 logit 模型条件似然估计的结果。与两期数据时的情况一样，我们看到成为母亲及在校生身份与较高的贫困风险相关，而与配偶同住及工作时间越长则与较低的风险联系在一起。

如何解释这些效应呢？拿 SPOUSE 的系数－0. 748 来说，取幂后我们得到的发生比为 0. 47。也就是说，如果一个女孩从没有与丈夫居住的状态变为与丈夫居住，那么她陷入贫困的几率要（在原来的基础上）乘以 0. 47。实际上就是，结婚将使少女陷入贫困的几率减少一半。而 HOURS 的系数－0. 0196 取幂之后得到发生比等于 0. 98。这是说，每个星期多工作一小时将使陷入贫困的几率缩小 2%。几个 YEAR 系数都是与年份 1 的对比，它们都是正的并且都在统计

上显著。注意，表中没有报告截距项，因为截距项已经“被条件出(is conditioned out)”似然函数。

表格中接下来的两列呈现的是采用广义估计方程法(GEE)估计 logit 模型得到的结果，它通过迭代广义最小二乘法(iterated generalized linear squares)修正了观察记录之间的相依问题。尽管结果的模式与条件似然分析的相似，但存在三个重要的差别。第一，SCHOOL 的系数从显著的正值变为不显著的负值。第二，GEE 分析中 MOTHER 和 SPOUSE 变量的系数明显要大得多，而几个年份变量的系数则都要小一些。第三，标准误都要更小。

表 3.4 的最后两列给出的是随机效应模型的最大似然估计。这一模型同样可以用方程 3.1 来表示，只是现在 α_i 被假定为一套随机变量，每一个都服从均值为 0 且方差恒定的分布，并且(最重要的是)与 x_{it} 保持相互独立。随机效应系数估计结果与 GEE 估计结果相似。所有的标准误都要比 GEE 估计的略大，但比条件似然分析的要小，因为条件似然法没有使用任何个体间的变异信息。事实上，条件似然法自动剔除了 324 个贫困状况在 5 年期间没有发生任何变化的少女(28%)。和两期数据时一样，如果一个人根本没有发生历时变化，那自变量也就没有什么可以解释的。

表 3.4 Logit 模型的条件似然估计及其他估计

	条件似然估计		GEE[a]		随机效应	
	系　数	标准误	系　数	标准误	系　数	标准误
MOTHER	0.582**	0.160	0.850**	0.092	1.077**	0.119
SPOUSE	−0.748**	0.175	−0.930**	0.121	−1.238**	0.152
SCHOOL	0.272*	0.113	−0.045	0.077	−0.064	0.098
HOURS	−0.0196**	0.0032	0.0209**	0.0023	−0.0267**	0.0029
YEAR2	0.332**	0.102	0.223**	0.073	0.287**	0.100
YEAR3	0.335**	0.108	0.171*	0.080	0.226*	0.104
YEAR4	0.433**	0.116	0.196*	0.084	0.256*	0.108
YEAR5	0.402**	0.127	0.122	0.093	0.172	0.115
截距			−0.543**	0.097	−0.681**	0.126

注：a. 同时设定了非结构化的相关矩阵，以及基于模型的标准误。

* $0.01 < p < 0.05$，** $p < 0.01$。

那么表 3.4 的三套结果中哪一套最好呢？三者最大的区别在于，GEE 和随机效应法都没有对未被观测的变量进行任何控制。相反，固定效应模型（条件似然法）控制了所有恒定变量，将每个女孩作为其自身的控制因素。而且其提供的标准误估计也是修正了相依问题的正确估计。不好的方面是：这些标准误要比随机效应模型及 GEE 估计的标准误大，因为数据集中有大量的信息没有被使用。权衡之后，对于这个例子，我更倾向于固定效应估计结果，因为它们受被忽略变量的影响而产生偏差的机会要小得多。不过，在个体内变异相对于个体间变异太小的情况下，固定效应系数的标准误可能会太大而无法容忍。

还有一点值得提醒的是，条件似然和随机效应估计都是“具体单位的（subject specific）”估计，而 GEE 估计只是“总体平均的（population averaged）”估计。那么二者的差别是什么呢？一个具体单位的系数能够告诉我们，如果一个具体个体的预测变量增加一个单位，那该个体会发生什么。而一个总体平均的系数只能告诉我们，如果每个人的预测变量都增加一个单位，那么整个总体会发生什么。如果模型是线性的，两种系数之间不存在差别。但对于 Logistic 回归模型，当然对于其他很多非线性模型也一样，具体单位系数一般要大于总体平均系数。

哪一个更好呢？答案取决于你的目的。如果你是一个医生，要估计某种他汀类药物能够在多大程度上降低你的病人罹患心脏病的风险，那么具体单位系数是很明显的选择。相反，如果你是一个政府卫生部门的官员，想知道如果风险人群中的每一个人都服用这种他汀类药物，因心脏病而死亡的人数将会有什么变化，那你最好使用总体平均系数。

当然，即使是在后一种公共卫生应用中，也有理由认为具体单位系数更具有实质性意义。假定真实模型是如方程 3.1 表达的基本随机效应 Logistic 模型。其回归系数向量 β 和 γ 二者都是具体单位的。如果我们用 xtlogit 命令通过 GEE 办法估计这一模型，我们将得到总体平均系数 β^* 和 γ^* 。这两套系数的差异程度取决于 α_i 的方差。特别是，如果 $\mathrm{var}(\alpha_i)=0$，那么 $\beta=\beta^*$，$\gamma=\gamma^*$ 。α_i 方差增加时，β^* 和 γ^* 的值将向 0 衰退。当 α_i 服从正态分布时，两者之间的近似关系为：

$$\beta^* \approx \frac{\beta}{\sqrt{0.346\text{var}(\alpha_i)+1}} \qquad [3.3]$$

因此总体均值系数取决于 Logistic 回归中未被观察的异质性的程度。对于上面的少女贫困问题数据，α_i 的方差估计值为 1.454。比较 GEE 系数估计值及随机效应系数估计值，我们发现上述关系确实近似地成立。

与时间的交互作用

条件似然法的另一个不足之处是它不能对非时变变量的回归系数进行估计(尽管这些变量都被潜在地控制住了)。不过，时变变量与非时变变量之间的交互项还是可以被放进模型中。表 3.5 中，模型 1 放入了一个变量，它是 MOTHER 与 BLACK 的乘积，其系数在 0.05 水平上显著。注意，与绝大多数含有交互项的模型不一样，这里根本不用(事实上，根本就不能)纳入 BLACK 的主效应。交互项解释起来和线性模型中的一样。MOTHER 的系数 0.982 代表着当 BLACK = 0 时，也就是说在非黑人少女中，MOTHER 的作用。取幂之后得到的对数发生比为 2.67。因此，对于非黑人少女来说，成为母亲将使陷入贫困的发生比在原来的基础上乘以 2.67。要得到成为母亲在黑人女孩中的影响，把主效应加上交互项系数，0.982－0.599 = 0.383，就能得到一个低得多的发生比 1.46。

在模型 2 中，我们可以看到 YEAR 与两个时变变量(SCHOOL 和 HOURS)还有两个非时变变量(BLACK 和 AGE)之间的交互作用显著。在这一模型中，YEAR 被当做一个定量变量而非定类变量，从而使模型及其解释得到简化[12]。YEAR 的编码取值为 0 到 4(而不是 1 到 5)，这样 SCHOOL 和 HOURS 的主效应可以被解释为 YEAR = 0 时这些变量的作用，也就说在第 1 个观察年份的作用。同在交互项中的 HOURS 和 AGE 则被表述为与它们各自对应均值的离差，这样可以使 YEAR 的主效应的解释变得容易。

表 3.5 带交互项的条件似然估计

	模型 1		模型 2	
	系 数	标准误	系 数	标准误
MOTHER	0.982**	0.253	0.687**	0.163
SPOUSE	−0.783**	0.178	−0.741**	0.178
SCHOOL	0.267*	0.113	−0.311	0.190
HOURS	−0.0192**	0.0032	−0.0060	0.0063
YEAR2	0.332**	0.102		
YEAR3	0.334**	0.108		
YEAR4	0.430**	0.117		
YEAR5	0.400**	0.128		
MOTHER* BLACK	−0.599*	0.290		
YEAR			0.021	0.059
YEAR* SCHOOL			0.251**	0.063
YEAR* HOURS			−0.0055*	0.0021
YEAR* BLACK			−0.181**	0.048
YEAR* AGE			−0.056*	0.023

注：* $0.01 < p < 0.05$，** $p < 0.01$。

对这些交互项的解释，时变变量与非时变变量之间有所不同。对于时变预测变量来说，通常最好是从各个自变量的作用是如何随时间变化而变化的角度来进行考虑。例如，SCHOOL 的作用可以表述为一个线性函数：$-0.311 + 0.251 \times \text{YEAR}$。因此，第 1 年时，它的作用是负的并且统计检验不显著。年份每增加 1，它的影响就增加 0.251，这样到第 5 年时它的作用就达到了 0.693（用发生比来说就是 2）。对于 HOURS 变量，其作用是 $-0.0060 - 0.0055 \times \text{YEAR}$。这样 HOURS 的作用最开始时是负的，并且随着年头的增加会持续地变得更负，到第 5 年时变为 −0.028。这相当于说每多工作 1 小时将使陷入贫困的发生比降低 2.8%。

对于非时变自变量，交互项最好的解释方式是查看 YEAR 的作用是如何随着这些变量的变化而变化的。根据这些变量的编码方式，YEAR 的主效应

(0.021)代表 YEAR 在那些既不是黑人也非在校生，并且年龄为平均开始年龄 15.65 岁，工作时间为平均工作时间 8.67 小时的少女中的作用。而在黑人少女中(其他特征都一样)，年份的作用是主效应加上交互作用 (0.021 − 1.81 = −0.16)。这相当于，每增加一年头，少女陷入贫困的发生比就减少 15%。我们也可以将 YEAR 的作用表达为第 1 个观察年份时年龄 AGE 的线性函数：0.021−0.056×(AGE−15.65)。如此，对 14 岁开始参与调查的此类少女来说，YEAR 的作用为 0.1134(大概是年份每增加 1 年，陷入贫困的发生比就增加 12%)；而对于 17 岁的此类少女，该作用为 −0.0546(从发生比上讲，大约每年下降 5%)。[13]

混合（模型）法

在第 2 章中我们将固定效应法和随机效应法统一到一个模型中。这是通过把时变自变量分解成个体内部分与个体间部分，然后用这两个部分一起拟合随机效应模型来实现的。个体间部分就是每个变量的个体均值(person-specific mean，即个体内均值)。个体内部分即(个体各观察值)与个体均值的离差。

现在我们将这种方法扩展到 Logistic 回归(Neuhaus & Kalbfleisch, 1998)。与在线性回归中一样，这种模型的魅力在于我们能够：(a)在模型中纳入非时变变量，(b)进行比较固定效应和随机效应的检验，以及(c)拟合更多类型的模型。(c)的一个例子是：与条件似然法不同，混合法能够使用其他的连接函数，如 probit 和互补双对数函数。

再一次使用 Stata 中的 xtlogit 命令，我对少女贫困问题这一数据拟合了随机效应模型，结果呈现在表 3.6 中。所有以 M 开头的变量名都是指个体均值。所有以 D 开头的变量名对应的都是相对于上述个体均值的离差。离差变量的回归系数从功能上讲与固定效应系数等价，因为估计时只利用了个体内的变异信息，从而控制了所有恒定变量。在第 2 章的线性混合模型中，离差变量的系数与最小二乘虚拟变量法产生的完全一样。但在这里，离差变量的系数与表 3.4 中的条件似然法系数并不相等，尽管它们确实很相近。

表 3.6 少女贫困数据的混合模型

	系 数	标准误	p
DMOTHER	0.594	0.158	0.000
DSPOUSE	−0.807	0.179	0.000
DSCHOOL	0.275	0.113	0.015
DHOURS	−0.0210	0.0032	0.000
MMOTHER	1.079	0.181	0.000
MSPOUSE	−2.146	0.255	0.000
MSCHOOL	−1.362	0.202	0.000
MHOURS	−0.0468	0.0058	0.000
BLACK	0.572	0.097	0.000
AGE	−0.123	0.050	0.013
YEAR2	0.333	0.101	0.001
YEAR3	0.330	0.107	0.002
YEAR4	0.431	0.115	0.000
YEAR5	0.391	0.125	0.002
截距	1.893	0.819	0.021

那些均值变量的系数本身并不是很有意思，引人注意的是它们(在量上)比对应的离差变量大多少。一个常规的随机效应模型(没有离析个体内与个体间成分时)内在地假定离差系数与均值系数相等。在混合模型中非常容易就能对这一假定进行检验，通过直接检验各对系数之间是否相等即可。表 3.7 清楚地表明有必要反对这一假设，这意味着固定效应模型在这里要比随机效应模型更好。在表 3.7 中，最重要的检验是联合检验，它检验的是四个离差系数同时都与对应的均值系数相等。[14]

表 3.7 对均值系数与离差系数是否相等的检验

	卡方值(Chi-square)	p
MOTHER	4.16	0.041
SPOUSE	19.31	0.000
SCHOOL	49.90	0.000
HOURS	15.70	0.000
联合检验(4 自由度)	79.10	0.000

混合方法的另一个优点，在于它能够得到非时变自变量的系数估计。例如，表 3.6 中，黑人有显著更高的贫困风险，而第一次访问时年龄更大的女孩的

贫困风险显著地低一些。不过，记住下面这点很重要：不像离差变量的系数，BLACK 和 AGE 的系数并没有控制住未被观测的自变量。

在第 2 章中，我们见识了混合线性模型能够扩展成为允许时变自变量带上随机系数的模型。这在混合 logit 模型中也是可能的，尽管估计此类模型在计算上非常精深。在 Stata 中，带有随机系数的 logit 模型需要使用一个不同的命令：xtmelogit（在 Stata10 中首先引进）。对这一少女贫困问题的例子，我估计了一个允许 DMOTHER 带有随机系数的混合模型。这一系数的估计均值为 0.603，标准差（不是标准误）为 0.751。这一标准差 95%的置信区间为 0.272 到 2.075。由于这一置信区间并不包含 0，因此，作为证据可以表明母亲身份对于贫困状况的影响在不同个人之间确实存在着变异性。

多分类反应变量的（固定效应）方法

到目前为止，我们只考虑了二分反应变量的情况。现在，有一个分类反应变量 y_{it}，它能够取两个以上的值。假设这些取值都是整数值，范围从 1 到 J，巡标（running index）为 j。令 $p_{itj} = \text{Prob}(y_{it} = j)$。接下来我们需要一个模型，以说明这些概率如何取决于预测变量x_{it}和z_{it}。

我们先从因变量的这些类别是序次排列的情况开始。针对序次分类因变量，最常用的模型是累积 logit 模型，也被称为序次 logit 模型。这一模型的固定效应形式可以如此表述：

$$\log\left[\frac{F_{ijt}}{1-F_{ijt}}\right] = \mu_{tj} + \beta x_{it} + \gamma z_i + \alpha_i,\ j = 1,\ \cdots,\ J-1 \qquad [3.4]$$

其中 $F_{ijt} = \sum_{m=j}^{J} p_{imt}$ 是落入某一个类别 j 或更高类别的"累积"概率。不幸的是，条件最大似然法不能用于这种模型，因它不能为 α_i 参数提供"简化充分统计量（reduced sufficient statistics）"。我们能做的是使用上一节讨论过的混合模型法，利用常规最大似然估计，配合稳健标准误，以调整各个体的多次观察之间的独立性不足。

作为示范，我们回到第 2 章中反社会行为的例子。在该数据集中，因变量

ANTI 是整数取值，并且取值范围为 0 到 6，但在线性回归模型中被当做定量因变量对待。这里，我们采取一种更恰当的方式把 ANTI 当做 Logistic 模型中的一个序次分类变量。

与二分情况一样，本混合模型法的实现，同样是先计算每个预测变量的分个体均值，然后计算(各个观测值)相对于这些均值的离差。再将均值变量及离差变量都纳入累积 logit 模型中，作为预测变量。如果想得到具体单位系数，我们就得估计随机效应模型，但是很难找到可以对序次 logit 模型进行这种估计的商业软件。无奈之下，我们只好进行常规最大似然估计，并采用稳健标准误，以修正重复观察之间的相依问题。

我是用 Stata 中的 ologit 命令完成上述任务的，结果在表 3.8 中。这些结果与表 2.8 中利用混合线性模型法得到的极其相似。所有的 P 值都能让我们对每个系数得出相同的结论。即便是系数本身及标准误也非常相近，当然在多数情况下并不会出现这样的结果。累积 logit 模型的系数表示落入因变量中较高类别而非较低类别的对数发生比的变化量。与二分类回归模型一样，对这些系数取指数幂之后就能得到发生比。要想知道更多有关如何解释这些系数的细节，可以参看笔者的另一本书(Allison，1999a)。

表 3.8 反社会行为的混合累积 logit 模型

	系 数	稳健标准误	p
DSELF	−0.064	0.013	0.000
DPOV	0.116	0.117	0.320
MSELF	−0.108	0.027	0.000
MPOV	0.696	0.185	0.000
BLACK	0.153	0.157	0.330
HISPANIC	−0.310	0.169	0.065
CHILDAGE	0.083	0.111	0.453
MARRIED	−0.189	0.163	0.247
GENDER	−0.598	0.128	0.000
MOMAGE	−0.017	0.029	0.557
MOMWORK	0.190	0.146	0.195
TIME_2	0.016	0.069	0.819
TIME_3	0.167	0.077	0.030

两个离差变量 DSELF 和 DPOV 的系数可以当作固定效应系数来解释。因为这些系数只取决于个体内的历时变化，并且它们控制了所有恒定的预测变量。对这两个离差系数等于对应均值系数的原假设进行检验(利用 Stata 中的 test 命令)，发现自由度为 2 时卡方值为 9.02，这在 0.01 水平上显著。与在线性模型中一样，其绝大部分原因来自于 MPOV 及 DPOV 系数之间太大的差异。这意味着即使通过稳健标准误修正了各观察之间的相依性，常规序次 logit 模型在这里仍不是合适的，至少对 POV 变量来说不合适。在此，我们得将注意力集中在离差系数上，因为它们控制了所有非时变自变量。

现在我们转到更复杂的情况，这里因变量各个类别之间并不存在序次关系。对于非序次分类变量使用最广的是多分类 logit 模型(multinomial logit model)，也被称为广义 logit 模型(generalized logit model)。下面是这一模型的固定效应形式：

$$\log\left(\frac{p_{ij}}{p_{iJ}}\right)=\mu_{ij}+\beta_j x_{it}+\gamma_j z_i+\alpha_{ij}\text{, } j=1\text{, }\cdots\text{, } J-1 \qquad [3.5]$$

公式 3.5① 可以被看作是一套联立的二分类 Logistic 回归方程，每个方程都将因变量的某一个类别与最后一个类别进行比较。固定效应 α_{ij} 在不同个体之间及不同反应值之间都可以变动，但不随时间变动。

固定效应多分类 logit 模型像二分类 logit 模型一样，可以为 α_{ij} 提供简化充分统计量，也即，为每个个体提供不同响应值的频次计数(frequency counts)。原则上讲，这一模型可以通过在限定那些计数的条件下，采用条件最大似然法进行估计(Chamberlain, 1980)。不过没有现成的商业软件可以实现这一点。如果时变预测变量是定类的，那么模型就可以转变为对数线性模型(log-linear model)并在相应的框架下进行估计(Conaway, 1989; Darroch & McCloud, 1986; Kenward & Jones, 1991; Tjur, 1982)。不过，建立一个那样的模型有一点复杂，这里我就不考虑该方案了。

另一种估计的方法是将多分类模型分解成为几个二分类模型，一个模型对应

① 原书为公式 3.4，根据上下文，当为编辑错误。——译者注

着一个特定类别与参照类之间的比较（Allison，1999a；Begg & Gray，1984）。然后，每个二分类模型都可以采用本章已经讨论过的条件 Logistic 回归方法进行估计。尽管这一方法可以产生近似无偏的系数估计，但估计结果将因参照类的选择不同而存在差异。另外，缺乏各个变量对因变量的作用的整体检验。

如我们在累积 logit 模型中看到的一样，要估计带有固定效应的多分类 logit 模型，混合模型法是最容易实现的方法。作为示例，我们回到本章大部分地方都在使用的例子，该数据中有 1151 名十几岁的少女，她们被连续观察 5 年，每年 1 次。不过这一次，我们将使用一个新的反应变量 EMPSTAT，它有如下 3 个类别：

(1) 正处于就业状态，

(2) 失业(下岗或正在找工作)，以及

(3) 退出劳动力市场(正在上学、在家料理家务等等)。

至于自变量，我们将使用 MOTHER（目前至少有一个孩子），SPOUSE（目前与丈夫同住），目前的年龄 AGE，以及 BLACK（与非黑人相对）。前 3 个变量为时变变量。

第一步是计算各个时变变量分个体的均值以及相对于这些均值的离差。由于有 241 条记录的反应变量 EMPSTAT 为缺失值，因此在删除了这些带有缺失值的记录之后再计算均值非常重要。

为了在 Stata 中估计多分类 logit 模型，我使用的是 mlogit 命令，并配以稳健标准误，来修正各个个体多次重复观察之间的相依问题。

结果呈现在表 3.9 的前两列数字中。从中可以看到两个二分类回归方程，每一个都将 EMPSTAT 中的一个类别与参照类——类别 1(处于就业中)——进行比较。这里的回归系数都是在假定所有观测彼此独立的条件下的常规最大似然估计，不过标准误却都修正了相依问题。需要记住的是，这些系数都是总体均值系数而非具体单位系数。

通过关注离差变量,我们可以看到成为母亲将增加失业或退出劳动力市场的发生比。与丈夫居住将提高失业(相对于就业)的发生比,但会减小退出劳动力市场(相对于就业)的发生比。当少女们年龄变大时,她们将稍微不那么容易失业,并且将相当地不那么容易退出劳动力市场。我还对每个离差变量系数是否等于对应的均值变量系数进行了检验,因此也就检验了固定效应模型与常规 Logistic 回归模型是否等价。对于这两个二分回归方程,卡方检验都高度显著。

表 3.9 就业状况的混合多分类 logit 模型

	GEE 估计(总体均值系数)				随机效应估计(具体单位系数)			
	失业 VS 就业		退出劳动力市场 VS 就业		失业 VS 就业		退出劳动力市场 VS 就业	
	系 数	稳健标准误	系 数	稳健标准误	系 数	标准误	系 数	标准误
DMOTHER	0.927**	0.160	0.799**	0.155	1.237**	0.201	0.951**	0.221
MMOTHER	1.656**	0.163	0.389*	0.167	2.185**	0.212	0.611**	0.234
DSPOUSE	0.640**	0.180	0.569*	0.221	0.816**	0.211	0.617*	0.269
MSPOUSE	0.678**	0.242	−1.111**	0.264	0.829**	0.305	−1.459**	0.394
DAGE	−0.070*	0.029	−0.381**	0.027	−0.131**	0.037	−0.503**	0.033
MAGE	−0.306**	0.045	−0.505**	0.046	−0.394**	0.061	−0.663**	0.062
BLACK	0.393**	0.096	0.499**	0.093	0.597**	0.127	0.720**	0.130
截距	4.381**	0.799	8.612**	0.802	5.630**	1.076	11.29**	1.091

注:* $0.01 < p < 0.05$,** $p < 0.01$。

在表 3.9 的最后两列中,我们可以看到另一套多分类 logit 模型估计结果,这是通过 Stata 的 xtlogit 命令估计两个分开的随机效应模型得到的。在第一个模型(失业相对于在职)中,所有“退出劳动力市场”的记录都被丢掉了。而第二个模型(退出劳动力市场相对于在职)中,所有处于“失业”类别的记录都被排除在外。比较随机效应估计值和 GEE 估计值,可以发现所有系数的符号和显著性水平差不多都相同。不过,随机效应估计结果在大小上普遍要大一些,因为它们是具体单位的而非总体平均的。

总结

第 2 章针对线性模型的所有固定效应方法都可以扩展应用到定类因变量上。基本要旨是一样的。固定效应方法能够控制所有未被观测的非时变变量的潜在干扰作用。不过在另一方面，相对于其他方法，固定效应方法趋向于具有较低的效率，因为个体间的变异未被考虑。当然，对于定类因变量，需要使用几个略有不同的估计程序。

本章的主要焦点在于二分类反应变量的回归模型。当每个个体恰好都只被观察两次时，固定效应 Logistic 模型可以用常规 Logistic 回归程序，通过条件最大似然法进行估计。这一方法需要如下几个步骤：放弃所有在两次观察中因变量取值相同的案例，将所有时变自变量都重新编码为差分值，然后对两个反应变量之一拟合常规二分类 Logistic 回归。

当每个个体的因变量仍然为二分变量，但都有两个以上的观察记录时，需要一种不同的数据结构，要求每个人的每一次回应都有一条单独的记录。但是，由于"简化充分统计量问题"，我们不能简单地估计含有标识的每个个体的虚拟变量的常规 logistic 回归。这种回归产生的估计系数将偏离 0，尤其是当每个个体的观察数很小时。解决之道在于使用条件最大似然法将固定效应移出似然方程。在 Stata 中，这可以通过 xtlogit 或 clogit 命令来完成。

很多研究者使用 GEE 估计或者随机效应 logistic 回归，而不是固定效应 logistic 回归，前两者都可以通过 xtlogit 命令进行。与固定效应方法相比，这两种方法都没有对未被测量的非时变解释变量进行任何控制。和固定效应一样，随机效应估计产生"具体单位"系数而不是"总体平均"系数。后者一般都会因为未被观测的异质性而逐渐向 0 消退变小。

固定效应和随机效应方法可以通过估计一个随机效应模型的方式综合成混合模型，估计前将时变预测变量分解为个体均值与相对这些均值的离差，然后再估计一个随机效应模型。如我们在第 2 章看到的一样，混合模型允许我们在其中纳入非时变变量，并且提供了一个比较固定效应模型和随机效应模型的

简单检验。

至于含有两个以上类别的反应变量，在商业软件中通常都无法对固定效应logistic回归进行条件最大似然估计。退而求其次，不管是针对序次因变量还是名义因变量，目前最好的办法还是使用混合模型法，并利用稳健标准误修正（各观察记录间的）相依问题。

第 4 章 | 计数变量的固定效应模型

我们的因变量经常会是某种计数(counts):小孩数、过去一年的性伴侣数、家里的电脑数量、过去 5 年被拘捕的次数等等。很多研究者把计数变量当作连续测量变量,并使用一般最小二乘回归进行分析。这样做可能是不恰当的,原因有一些。例如,计数变量必定是离散型的,并且取值不能小于 0。它们的分布通常都是高度偏态的。

一种通常更好的办法是估计泊松回归模型(poisson regression model)或负二项回归模型(negative binomial regression model),这两种方法专门被设计用来对计数变量建模(Long, 1997)。在对它们进行简略介绍后,我们将考察如何扩展这些计数变量模型,以处理每个个体被观察多期的数据,并带上固定效应以控制所有非时变预测变量。[15]在此过程中,我们将再次遇到上一章讨论二分类结果变量模型时出现的很多问题。不过,困扰 logistic 回归的那些估计问题在计数变量模型中没有那么严重。

我们先来考虑一下将会贯穿本章的例子。这一数据包含 346 个制造业公司(manufacturing firm),记录了从 1975 年到 1979 年每个公司每年获得的专利数量。有关这一数据的已有分析可以在霍尔、格瑞里奇及豪斯曼的著作(Hall, Grilliches & Hausman, 1986)以及卡梅伦和特里维迪的著作(Cameron & Trivedi, 1998)中找到。原始数据集中每个公司有一条记录,每条记录中变量 PAT75 到 PAT79,包含这 5 年各年的专利数量。作为预测变量,我们有 1970 年到 1979 年每个公司研究与发展开支(research and development expenditures)的对数值(LOGR70 到 LOGR79)。同时也包括两个非时变变量:LOGSIZE 是 1972 年公司账面价值(book value)的对数;SCIENCE 是虚拟变量,如果公司属

于科技行业(science sector)则为 1,不属于科技行业则为 0。

每个个体被观察两期的计数数据泊松模型

和在前面几章看到的一样,当每个个体只有两期观察数据时,线性及 logistic 固定效应分析可以通过常规软件用简化方法完成。对于计数数据,这同样可行。事实上,固定效应泊松回归模型可以采用用于分组数据(grouped data)的普通 logistic 回归程序进行估计。

为了用专利数据说明这一点,我们将忽略中间的年份,而只关注 1975 年和 1979 年的记录。令 y_{i1}表示公司 i 在 1975 年的专利数,y_{i2}表示在 1979 年的专利数。这两个变量都被假定服从期望值为 λ_{it}的泊松分布。这样,$y_{it}=r$的概率由如下方程给定:

$$\Pr(y_{it}=r)=\frac{\lambda_{it}^{r}e^{-\lambda_{it}}}{r!},\ r=0,\ 1,\ 2,\ \cdots \qquad [4.1]$$

泊松分布可能是所有分布中最简单的适合计数数据的概率分布。它可以从满足如下假定的随机过程模型中推导出来:(a)事件(在这里是专利)不能同时发生,并且(b)事件之间彼此独立(Cameron & Trivedi, 1998)。所谓彼此独立,意思是一个事件的发生并不会提高或降低将来事件的发生概率。

注意,我们的模型并没有假定整个样本都服从某一单一泊松分布。相反,每个公司的专利数都来源于不同的泊松分布,分布的期望值 λ_{it}在不同公司及不同时期都可以不同。泊松分布的独特性在于它的均值和方差相等:

$$E(y_{it})=\mathrm{var}(y_{it})=\lambda_{it} \qquad [4.2]$$

不幸的是,我们将看到,这一性质有时也会导致一种被称为过离散(overdispersion)的问题,这种问题会严重危害泊松回归模型的估计。

接下来,我们令 λ_{it}作为自变量的对数线性函数(log-linear function)

$$\log\lambda_{it}=\mu_{t}+\beta x_{it}+\gamma z_{i}+\alpha_{i} \qquad [4.3]$$

与前面各章一样,x_{it}表示时变预测变量,z_i 则表示非时变预测变量,α_i 表示

未被观察的“固定效应”。和以前一样，把 α_i 当作一套固定的常数，等价于将它们当作是可以与x_{it}存在任意的不受限的相关的随机变量。向量x_{it}包括当前年份 t 及此前 5 年每一年的研究与发展开支。

我们的目标是估计出方程 4.3 中的参数。为了实现这一点，需要使用条件最大似然法，即第 3 章中用来估计固定效应 logistic 模型的方法。考虑到 y_{i2}的分布以两个时期汇总的事件总数(表达为 $w_i = y_{i1} + y_{i2}$)为条件，因此它可以表示为 $y_{i2} | w_i — B(p_i, w_i)$。也就是说，在满足总数量为 w_i 的前提下，1979 年的专利数 y_{i2}服从参数为 p_i 和 w_i 的二项分布，其中：

$$p_i = \frac{\lambda_{i2}}{\lambda_{i2} + \lambda_{i1}} \qquad [4.4]$$

经过一些运算，得到：

$$\log\left(\frac{p_i}{1-p_i}\right) = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) \qquad [4.5]$$

这样，我们就已将泊松回归模型转变为自变量为原始自变量差分值的 logistic 回归模型。注意，和以往的情况一样，α_i 和 γz_i 都被从方程 4.5 中消除掉了。

为了在 Stata 中实现这一条件法，我使用了 blogit 命令，它可以对分组二项数据(grouped binomial data)进行最大似然估计(ML estimation)。blogit 命令要求因变量包括两个部分：“事件(events)”数以及“试验(trials)”数。通过提交下面的命令，我首先估计了一个不带任何自变量(仅含截距项)的模型。

```
blogit pat79 total
```

其中 PAT79 是 1979 年的专利数，TOTAL 等于 PAT75＋PAT79。估计得到的截距为 －0.1386，对应的标准误为 0.0129，所得 z 统计量为 －10.68。这告诉了我们什么呢？如果用 m_1 表示年份 1 的平均专利数，而 m_2 为年份 2 的平均专利数的话，这一截距其实就是 $\log(m_1/m_2)$。如果这两个年份的专利数恰好一样，那截距就会等于 0。上述结果为负，表明平均专利数随着时间推延在下

降。更具体的说，如果我们计算：

$$100[\exp(-0.1386)-1]=-12.9\%$$

就可以得到平均值从 1975 年到 1979 年下降的比例。另外，由于截距对应的 z 统计量如此之大，我们可以拒绝认为这两个年份的均值相等的虚无假设。

事实上，这个 z 统计量太大了。由于所谓的过离散问题，基于泊松分布估计所得的常规标准误只是真实标准误的一个低估值(underestimate)。在后文我们将更详细地讨论离散问题。在那以前，我们可以在 Stata 中使用刀切法(jackknife)或者自助法(bootstrap)选项，以获得稍微好一些的标准误估计。这些计算密集型方法(computer-intensive methods)利用数据集的众多子样本或再抽样样本反复进行分析，以估计标准误(更详细的解释可以看 Mooney & Duval, 1993)。这里，刀切法标准误为 0.0371，产生的 z 统计量为 -3.74。自助标准误为 0.0358，对应 z 统计量为 -3.78。尽管这些 z 统计量比原来的常规 z 统计量要小很多，但很显然，它们仍然是高度显著的。

下一步将纳入自变量，它们是研究与发展经费支出对数的差分值。为了与以往对这一数据的分析保持一致，我们的分析目标是同时纳入“当前(current)”的研究与发展开支和此前 5 年每 1 年开支的时滞值(lagged values)。为此，我定义了如下变量：

$$\begin{aligned}
RD0 &= LOGR79 - LOGR75\\
RD1 &= LOGR78 - LOGR74\\
RD2 &= LOGR77 - LOGR73\\
RD3 &= LOGR76 - LOGR72\\
RD4 &= LOGR75 - LOGR71\\
RD5 &= LOGR74 - LOGR70
\end{aligned}$$

RD0 是计算专利数那两年的差分值，RD1 到 RD5 是 1 到 5 年的时滞值的差分值(difference scores for lags of 1 to 5 years)。这 6 个变量都被纳入分组

logistic 回归模型作为解释变量，结果呈现在表 4.1 的模型 1 中。

检查这些参数的估计值及相应的统计量，可以看到 RD0，研究与发展支出的同期测量值对专利数有着高度显著的影响，系数为 0.5214。解释这一系数时，得留意因变量（专利数的期望值）和自变量（研究与发展开支）都被取了对数（见方程 4.3）。因为两个变量都被取了对数，我们可以说，在控制了研究与发展经费的时滞测量后，研究发展经费 1%的增长与同一年期望专利数 0.52%的增长有关。开支的时滞测量值的影响比这要小得多。

这里我们仍需要使用更为稳健的标准误估计以处理过离散问题。呈现在表 4.1 中的自助标准误达到常规标准误的两倍。使用自助标准误后，我们发现只有 RD0 仍保持统计显著，而且即使是这个变量，它的 z 统计量也大大地变小了。

表 4.1　专利数据的条件泊松估计——两个时期

	模型 1			模型 2		
	系　数	常规标准误	自助标准误	系　数	常规标准误	自助标准误
RD0	0.521	0.084**	0.207*	0.533	0.085**	0.209*
RD1	−0.207	0.113	0.227	−0.192	0.113	0.256
RD2	−0.118	0.111	0.277	−0.137	0.111	0.341
RD3	0.060	0.096	0.263	0.062	0.096	0.314
RD4	0.181	0.090*	0.244	0.183	0.091*	0.209
RD5	−0.093	0.069	0.118	−0.100	0.069	0.167
SCIENCE				0.023	0.028	0.089
LOGSIZE				0.017	0.008*	0.017
截距	−0.222	0.018**	0.052**	−0.347	0.062**	0.138*

注：* $0.01 < p < 0.05$，** $p < 0.01$。

和我们前面的固定效应模型一样，表 4.1 中的估计也都控制住了所有在不同时期保持稳定不变的变量。尽管系数解释起来不那么直截了当，我们仍可以把不随时间变化而变化的自变量放入模型。表 4.1 中的模型 2 含有虚拟变量 SCIENCE（是否属于科技行业）以及 LOGSIZE（公司账面价值）。当我们使用自

助标准误时,这两个变量都没有达到统计显著。它们的系数可以解释为与时间的交互作用的大小。与所有的交互项一样,这些系数可以采用两种不同的解释方式。例如,SCIENCE 的系数 0.0275 可以表示 SCIENCE 在 1979 年的系数与在 1975 年的系数的差异。它在统计上高度不显著,说明这一变量在这两个年份有着相等的作用。或者,我们可以将 0.0275 解释为时间对属于科技行业的公司的影响的增加量,相对于那些不属于科技行业的公司。当然,因为它远没有达到统计显著的水平,我们可以得出如下结论:这两种不同行业的公司的专利数的变化率实质上是一样的。类似的解释也可对 LOGSIZE 做出。

多期数据泊松模型

当个体在两个以上的时期得到观察时,固定效应泊松模型的估计需要采取不同的方法。现在我们继续上文的例子,分析 1975 年到 1979 年每一年的专利数——用 y_{it}表示。和前面一样,我们假定每个 y_{it}都来自公式 4.1 给定的期望值为 λ_{it}的泊松分布,令 λ_{it}像方程 4.3 所给定的那样,是自变量的对数线性函数。

有两种方法对这一模型进行估计,条件最大似然估计和无条件最大似然估计。在条件最大似然估计中,似然函数建立在每个个体(不同时期)的专利数的总和一定的条件上,它能消除固定效应(α_i)。所得条件似然值(Cameron & Trivedi, 1998)与下面的式子成比例:

$$\prod_i \prod_t \left[\frac{\exp(\mu_i + \beta x_{it})}{\sum_s \exp(\mu_s + \beta x_{is})} \right]^{y_{it}} \qquad [4.6]$$

在 Stata 中,这一似然值可用 xtpoisson 命令来进行最大化(这一命令同样可估计随机效应及总体平均模型)。该命令要求数据集已被重构为每个公司每一年都有一条记录的形式,并且有一个共同的 ID 变量将来自同一公司的 5 条记录连在一起[16]。新数据集有来自 346 个公司的 1730 条记录。表 4.2 展示了样本前 4 个公司的 20 条记录。

表 4.2　重构数据集前 4 个公司的观察记录

观测值	ID	t	PATENT	RD0	RD1	RD2	RD3	RD4	RD5
1	1	1	32	0.92327	1.02901	1.06678	0.94196	0.88311	0.99684
2	1	2	41	1.02309	0.92327	1.02901	1.06678	0.94196	0.88311
3	1	3	60	0.97240	1.02309	0.92327	1.02901	1.06678	0.94196
4	1	4	57	1.09500	0.97240	1.02309	0.92327	1.02901	1.06678
5	1	5	77	1.07624	1.09500	0.097240	1.02309	0.92327	1.02901
6	2	1	3	−1.48519	−0.68464	−0.15087	0.08434	−0.21637	−0.45815
7	2	2	2	−1.19495	−1.48519	−0.68464	−0.15087	0.08434	−0.21637
8	2	3	1	−0.60968	−1.19495	−1.48519	−0.68464	−0.15087	0.08434
9	2	4	1	−0.58082	−0.60968	−1.19495	−1.48519	−0.68464	−0.15087
10	2	5	1	−0.60915	−0.58082	−0.60968	−1.19495	−1.48519	−0.68464
11	3	1	49	3.67434	3.58542	3.52962	3.44199	3.40697	3.39054
12	3	2	42	3.77871	3.67434	3.58542	3.52962	3.44199	3.40697
13	3	3	63	3.82205	3.77871	3.67434	3.58542	3.52962	3.44199
14	3	4	77	3.88021	3.82205	3.77871	3.67434	3.58542	3.52962
15	3	5	80	3.90665	3.88021	3.82205	3.77871	3.67434	3.58542
16	4	1	0	0.43436	0.53714	0.53714	0.58779	0.48454	0.54340
17	4	2	0	0.33836	0.43436	0.43436	0.53714	0.58779	0.48454
18	4	3	1	0.36561	0.33836	0.33836	0.43436	0.53714	0.58779
19	4	4	0	0.43860	0.36561	0.36561	0.33836	0.43436	0.53714
20	4	5	0	0.42459	0.43860	0.36561	0.36561	0.33836	0.43436

和在两期数据情况下一样，我们的回归模型中包括当年的研究发展经费及此前5年的经费。模型也纳入了与5个年份中的4个对应的年份虚拟变量组（第1年被作为参照类）。表4.3中的结果与我们表4.1中用5年中2个年份数据得到的相似。即当年的研究发展经费支出（RD0）有很强的作用，而时滞值影响要弱得多（RD1到RD5）。TIME系数表明：在这5年期间专利数有显著的下降趋势。请注意：固定效应模型没有报告截距信息，因为截距项被从条件似然函数中消除了。

在该表中，分别用常规方法及自助法对固定效应（条件似然法）标准误进行了估计。[17]和两期分析一样，自助标准误要比常规标准误大得多，多数情况下前者都接近后者的两倍。原因还是泊松回归中非常常见的过离散问题。大体上讲，过离散意味着事件计数的实际变异要比基于某一泊松分布的期望变异多。这是经常发生的，因为回归模型通常难以囊括解释这些计数变异的所有原因。不过，由于我们估计的是固定效应模型，已经控制住了公司间在专利数上的所有变异。因此，能够引发过离散问题的被忽略变量只能是那些随着时间变化在公司内有所变动的变量。在含泊松回归的某些软件（如SAS）中，你可以得到某种被称为偏差（deviance）的统计量，它能够直接测量过离散的程度。但Stata不为常规泊松回归报告偏差统计量，因此使用自助标准误或刀切标准误以避免潜在错误总是一种好办法。

为了进行比较，表4.3同时报告了使用xtpiosson估计的另外两种模型，随机效应模型和总体均值模型（由广义估计方程或GEE估计所得）。与固定效应模型一样，随机效应模型也能用方程4.1和4.3描述，只是其中的α_i被假定为具有特定概率分布的随机变量，并且和x_{it}及z_i相互独立。这一独立假定，意味着随机效应模型没有控制未被观察的协变因素（unobserved covariates）。

Stata中的默认设置假定α_i服从对数伽马分布（log-gamma distribution），不过也能将其设定为正态分布。而在总体均值模型中，并没有假定泊松回归方程中还存在一个干扰项，而只是允许每个公司的各次观察之间存在相关。[18]这种总体均值模型是通过GEE方法进行估计的，如在logistic模型中一样，这种

方法是一种迭代广义最小二乘法(iterated generalized least squares)。随机效应及GEE估计都很容易受过离散问题影响,因此常规标准误是有偏的。针对随机效应模型,我报告了自助标准误。对于GEE模型,我报告了更容易计算的稳健标准误,但xtpoisson命令中无法为随机效应及固定效应模型提供这一标准误。

表4.3 专利数据的泊松回归估计——5个时期

	固定效应			随机效应		GEE估计	
	系 数	常规标准误	自助标准误	系 数	自助标准误	系 数	稳健标准误
RD0	0.322	0.046**	0.084**	0.477	0.072**	0.303	0.053**
RD1	−0.087	0.049	0.087	−0.008	0.058	0.049	0.056
RD2	0.079	0.045	0.064	0.136	0.061*	0.167	0.051**
RD3	0.001	0.041	0.072	0.059	0.090	0.085	0.062
RD4	−0.005	0.038	0.065	0.028	0.051	0.050	0.042
RD5	0.003	0.032	0.063	0.082	0.067	0.038	0.043
TIME 2	−0.043	0.013**	0.017*	−0.047	0.016**	−0.048	0.017**
TIME 3	−0.040	0.013**	0.026	−0.056	0.024*	−0.052	0.026*
TIME 4	−0.157	0.014**	0.036**	−0.190	0.041**	−0.178	0.043**
TIME 5	−0.198	0.015**	0.033**	−0.253	0.038**	−0.234	0.041**
截距				−1.403	0.081**	1.828	0.123**

注:* $0.01 < p < 0.05$, ** $p < 0.01$。

与我们在此前的比较中看到的一样,固定效应估计的标准误要比随机效应及GEE模型的大。一如平常,这是因为固定效应只使用了公司内的变异而完全放弃了公司间变异的原因。事实上,5年期间每年的专利数为0的公司根本就被排除在条件似然函数之外。这一数据集中了22个这样的公司。从好的方面看,固定效应估计控制了所有稳定的公司属性,而随机效应及GEE估计只控制了被明确纳入模型的那些公司层面的性质(在这几个模型中没有纳入一个这样的变量)。就本分析而言,三种方法的结果唯一的主要区别在于:随机效应及GEE模型结果显示RD2有一定作用,而固定效应模型中并不存在此类证据。

固定效应泊松回归模型也可以使用无条件最大似然法进行估计。这是通

过估计一个含有标识所有公司(少一个)的虚拟变量组的常规泊松模型来实现的。在讨论 logistic 回归模型的第 3 章中,我们已经看到条件最大似然估计和无条件最大似然估计产生了不同的估计结果。而且,无条件最大似然估计是错误的——他们倾向于产生太大的系数估计。不过,在泊松回归模型中,条件及非条件最大似然估计总是产生相同的结果(Cameron & Trivedi, 1998)。因此,选择哪一个纯粹是看哪个计算起来方便。在 Stata 中,使用无条件法分析本专利数据所花的时间要长得多,因为需要估计超过 300 个虚拟变量的系数。不过,很多软件包(如 SAS)不含进行条件泊松回归的程序,在这种情况下就只好选择无条件最大似然估计了。

表 4.3 中的自变量都是时变变量。我们是否也能在固定效应模型中纳入非时变变量呢? 在上文中,当每个公司只有两条记录时,我们在用于泊松模型条件估计的 logistic 模型中纳入了两个非时变变量。而这些变量的系数被解释为与时间的交互作用。但是此刻,非时变变量不能被直接纳入模型。不过,我们可以设置非时变变量与时变变量,包括时间本身之间的交互项。例如,有人可能假定研究发展经费支出对科技公司专利数的影响要比在非科技公司的大。表 4.4 报告的是一个纳入了 SCIENCE 和 RD0 乘积项的模型的结果。可以看到,没有必要(甚至根本不能)纳入 SCIENCE 的主效应。简单起见,这一模型中删除了在表 4.3 中不显著的研究发展开支的时滞效应。

表 4.4 含有非时变协变量的条件泊松估计

	固定效应		
	系 数	常规标准误	自助标准误
RD0	0.375	0.048**	0.078**
RD0 * SCIENCE	−0.204	0.067**	0.188
TIME_2	−0.034	0.013**	0.014*
TIME_3	−0.034	0.013**	0.020
TIME_4	−0.151	0.014**	0.031**
TIME_5	−0.189	0.015**	0.035**

注: * $0.01 < p < 0.05$, ** $p < 0.01$。

从表 4.4 中,我们可以看到,在使用常规标准误时,RD0 与虚拟变量 SCIENCE 之间的交互作用显著,而在自助标准误下,这一交互作用并不显著。但不管是在哪种情况下,交互项的作用都与假设——研究发展经费对专利数的影响在科技公司要比非科技公司大——相反。更具体地说,研究发展经费在非科技公司的影响是 RD0 的主效应,即 0.375。而在科技公司,其作用是 0.375 − 0.204 = 0.171,等于主效应加上交互作用。

现在我们检验一下专利数的变化速率在科技行业与非科技行业是否存在差异。对于表 4.5 中的模型,我限定时间为线性作用,然后纳入一个 SCIENCE 与 TIME 的交互作用项。表 4.5 中的结果未能证明科技公司和非科技公司的专利数在变化速率上有所不同。交互项系数还远没有达到统计显著的水平(不管是用常规标准误还是自助标准误),其大小也仅仅只有时间主效应的 2%。

表 4.5 条件泊松估计——含与时间的交互作用

	固定效应		
	系 数	常规标准误	自助标准误
RD0	0.276	0.039	0.075
TIME	−0.049	0.005	0.010
SCIENCE * TIME	−0.001	0.006	0.016

计数数据的固定效应负二项模型

如我们刚刚已经看到的,固定效应泊松回归模型很容易受到过离散效应的影响。这多少有些出人意料,因为固定效应模型已经通过 α_i 参数允许不同个体之间存在未被观测到的异质性。但是这种异质性被假定不随时间变化而变化,然后仍然可能存在着仅仅属于某些特定时点的未被观测到的异质性,正是它们导致了观察到的过离散。正如我们所见的,在过离散情况下,可以通过使用自助法及刀切法对标准误进行矫正。尽管这种方法并不坏,但通过直接将过离散问题建构到事件计数模型中,我们可能做得更好。

为了模拟这种过离散，我们假定每个公司每个时点的专利数来自于某一个负二项分布。负二项分布是一般化的泊松分布，通过一个额外的参数，它允许过离散存在。负二项模型的吸引之处在于，它所估计的系数更加有效（抽样变异更小），而且其标准误及统计检验比诸如自助法及刀切法之类的经验的、事后的调整更精确。

不过，负二项回归模型公式化的方式不止一种。这里使用的是被卡梅伦和特里维迪（Cameron and Trivedi，1989）叫做NB2的模型，其中 y_{it} 的概率质量函数（pmf）是这样定义的：

$$\Pr(y_{it}=r)=\frac{\Gamma(\theta+r)}{\Gamma(\theta)\Gamma(r+1)}\left(\frac{\lambda_{it}}{\lambda_{it}+\theta}\right)^{r}\left(\frac{\theta}{\lambda_{it}+\theta}\right)^{\theta} \qquad [4.7]$$

在这一等式中 λ_{it} 是 y_{it} 的期望值，θ 是过离散参数，$\Gamma(\cdot)$ 是伽马函数。当 $\theta\to\infty$，这一分布向泊松分布收敛。如在泊松模型中一样，我们假定 y_{it} 的期望值可以被一个对数线性方程描述：

$$\log\lambda_{it}=\mu_t+\beta x_{it}+\gamma z_i+\alpha_i \qquad [4.8]$$

其中 α_i 作为固定效应处理。在控制 α_i 的条件下（conditional on α_i），每个个体（如本例题中的一个公司）的几个计数值被假定彼此独立，尽管在无条件状况下，它们可能互相依赖。

这一模型如何能被估计出来呢？不像泊松模型，在这里不能用条件似然法。用技术术语来说，每个个体的计数总和并不是 α_i 的“完全充分统计量（complete sufficient statistic）”，因此以总数为条件（conditioning on the total counts）并不能将 α_i 从似然方程中消除掉。豪斯曼、霍尔及格瑞里奇（Hausman，Hall & Griliches，1984）提出了一种非常不同的固定效应负二项回归模型，他们为该模型推导出一种条件最大似然估计量。事实上，他们的方法已经被吸收进Stata的xtnbreg命令中。不过埃里森和沃特曼（Allison & Waterman，2002）已表明这种方法并非真正的固定效应回归模型，它事实上并没有控制所有的固定自变量，下面我们就会看到这一点。

相反，我们将进行无条件最大似然估计，通过估计纳入了标识所有个体（除

其中一个外)的虚拟变量的负二项回归模型的方式。在 Stata 中,这可以通过 nbreg 命令实现。[19]这一模型的运算非常慢,因为有大量公司虚拟变量的系数需要估计。为了稍微提高速度,我忽略掉 22 个在此 5 年中没有任何专利的公司。这些公司对似然函数没有任何贡献,并且其虚拟变量系数不收敛。

表 4.6 中的结果应该与表 4.3 中固定效应泊松回归的结果进行比较。这里没有将与各个公司对应的虚拟变量的系数呈现出来。显然,负二项回归模型的系数与泊松模型的结果非常相似。而且,负二项模型的标准误及检验统计量与自助标准误泊松模型的相近。标签为阿尔法(Alpha)的估计参数是离散性的一种测量。事实上,它是 $1/\theta$ 的估计,其中 θ 为方程 4.7 中的那个参数。很明显阿尔法大于 0,系数取值达到其对应标准误取值的 10 多倍,这意味着存在显著的过离散。

表 4.6　固定效应负二项模型的无条件估计

	固定效应模型		
	系　数	常规标准误	梯度外积(OPG)标准误
RD0	0.371**	0.063	0.072
RD1	−0.083	0.068	0.073
RD2	0.064	0.064	0.075
RD3	0.014	0.060	0.071
RD4	0.034	0.056	0.060
RD5	0.002	0.046	0.052
TIME_2	−0.049*	0.023	0.027
TIME_3	−0.051*	0.023	0.029
TIME_4	−0.159**	0.024	0.028
TIME_5	−0.224**	0.025	0.028
截距	3.677	0.118	0.101
Alpha	0.020**	0.002	0.002

注: * $0.01 < p < 0.05$, ** $p < 0.01$。

Stata 还报告了 Alpha = 0 这一虚无假设的似然比卡方统计量,在这里它的取值为 499.54,对应自由度为 1,无论以什么标准来说,这都是统计显著的。这一统计量是通过将负二项模型的对数似然值(log-likelihood)与泊松模型对数似然值的差值乘以 2 计算得到的。这样做可行是因为:泊松模型是负二项模型在 Alpha 值

等于 0 时的特例。这一检验意味着我们应该拒绝泊松模型，选择负二项模型。

显然，负二项模型对这一数据的拟合要比泊松模型好得多。与泊松模型（条件估计与无条件估计必然相同）不同，无条件负二项估计无法保证能够抵抗来自由伴随性参数问题（在第 3 章 logistic 模型中讨论过）造成的偏差。通过使用蒙特卡罗模拟（Monte Carlo simulation），埃里森和沃特曼（Allison & Waterman，2002）发现：无条件负二项回归估计量没有呈现任何相对于伴随性参数的真正偏差。他们同样表明：负二项估计量有着比泊松估计量确实要小的标准误。不过，无条件的负二项估计的确有一个缺陷：其置信区间倾向于太小（尽管总体差值[undercoverage]远非泊松模型那么严重）。在很多情况下，名义的 95%的置信区间只有 85%的机会覆盖真值。这一问题可以通过对过离散产生的标准误进行调整而轻松地解决，调整所使用的是一个基于离差统计量的公式。在模拟中经此调整以后，几乎在所有情况下，实际覆盖率都已非常接近名义的 95%的置信区间。尽管 Stata 不报告这一矫正所需的离差统计，我发现：由 vce(opg)选项产生的标准误与由离差矫正产生的标准误一样。这些标准误呈现在表 4.6 的第 3 列。

对于专利数据的例子来说，负二项估计的运算时间还能够容忍，但对于非常庞大的数据集来说就会是很大的问题，届时将有大量虚拟变量的系数需要估计。Greene(2001)给出了此种运算难题如何迅速得到解决的办法，但是需要对现有 Stata 计算法则进行调整。

在前面，我曾提到 Stata 的 xtnbreg 命令下的条件负二项法并非真正的固定效应法。表 4.7 为这一事实提供了说明。这些估计结果是用 xtnbreg 命令及固定效应选项产生的。在模型 1 中我们只纳入了时变预测变量，结果与我们在表 4.6 中看到的非常相似。但我们得到了截距项的估计值，这本该已经在条件似然函数中消除了。模型 2 包括两个非时变自变量，SCIENCE 和 LOGSIZE。如果条件似然法真的控制了所有非时变变量，那我们应该无法纳入这些变量，因为它们是冗余的。另外，我们发现 LOGSIZE 有高度显著的系数，而 RD0 的作用因为 SCIENCE、LOGSIZE 的纳入而发生变化。这些对于真正的固定效应估计量来说都是不合常理的。

表 4.7 Stata“固定效应”负二项模型的估计结果

	模型 1		模型 2	
RD0	0.319**	0.067	0.273**	0.071
RD1	−0.080	0.077	−0.098	0.077
RD2	0.056	0.071	0.032	0.071
RD3	−0.013	0.066	−0.020	0.066
RD4	0.035	0.062	0.016	0.063
RD5	0.009	0.052	−0.010	0.053
TIME_2	−0.042	0.025	−0.038**	0.024
TIME_3	−0.049	0.025	−0.040**	0.025
TIME_4	−0.161**	0.026	−0.144**	0.026
TIME_5	−0.215**	0.026	−0.196**	0.027
SCIENCE			0.018	0.198
LOGSIZE			0.207**	0.078
截距	2.424**	0.175	1.661**	0.343

注：** $p < 0.01$。

混合（模型）法

如在前文看到的一样，有可能将固定效应与随机效应方法结合起来，从而获得各自的一些优点。在这一框架内，我们可以进行统计检验，以比较固定效应模型和随机效应模型，还可以对不随时间变化而变化的变量的影响进行估计。一如从前，第一步是为各个体计算每个时变自变量的均值，然后计算相对于这些均值的离差值。接下来的步骤是运行一个同时将离差变量和均值变量作为自变量的回归模型。在这里，我们将估计一个负二项回归模型，因为它们不那么容易犯过离散问题。为了得到正确的标准误，所用的估计方法是否能允许各个个体的多次观察之间存在相依性将会非常重要。随机效应模型和总体均值(GEE)模型都能做到这一点。[20]

利用 Stata 的 xtnbreg 命令，我把随机效应模型和 GEE(总体均值)模型都估计出来了，结果呈现在表 4.8 中。对于 GEE 模型，我估计的是默认的“可交换(exchangeable)”模型，这种模型假定同一个公司所有年份之间的相关相等，这

使得其在本质上与随机效应模型等价。所有以字母 D 开头的变量名代表的是离差变量，而以 M 开头的变量代表分企业的均值。

表 4.8　负二项回归的混合模型估计

	随机效应		广义估计方程(GEE)	
DRD0	0.322**	0.071	0.410**	0.120
DRD1	−0.057	0.076	−0.129	0.120
DRD2	0.081	0.068	0.056	0.082
DRD3	−0.006	0.064	−0.012	0.095
DRD4	0.011	0.059	0.007	0.099
DRD5	0.019	0.050	−0.062	0.088
MRD0	−0.336	0.697	0.031	0.798
MRD1	2.246	1.426	1.080	1.722
MRD2	−1.985	1.585	−1.110	1.850
MRD3	−0.500	1.408	−0.075	1.566
MRD4	1.248	1.106	1.119	1.136
MRD5	−0.051	0.517	−0.274	0.478
SCIENCE	0.057	0.103	−0.007	0.112
LOGSIZE	0.119**	0.045	0.105*	0.052
TIME_2	−0.042*	0.021	−0.052	0.034
TIME_3	−0.049*	0.022	−0.049	0.040
TIME_4	−0.168**	0.023	−0.100*	0.047
TIME_5	−0.208**	0.025	−0.209**	0.050
截距	1.038**	0.171	1.002	0.178

注：* $0.01 < p < 0.05$，** $p < 0.01$。

离差变量的系数可以当作固定效应估计值来解释，因为其仅仅基于公司内的变异，也正因为如此，它们控制了所有固定的预测变量。事实上，它们非常接近表 4.6 中研究发展经费变量的固定效应系数。和那个表格一样，唯一达到统计显著的离差变量是 DRD0，即当年研究发展经费支出的对数。GEE 系数 0.41 说明研究发展经费增长 1%与专利数增加 0.41%相关。

与平常一样，混合模型法的一个吸引力在于它能够纳入非时变预测变量，在这里是 SCIENCE 和 LOGSIZE。后者对于专利数有着显著的正影响。不过要谨记，与离差变量的系数不同，这些系数并没有控制其他未被纳入的解释变量。

混合模型的另一魅力在于它具备对固定效应模型与限制更多(more restricted)的随机效应模型进行比较检验的能力。这是通过检验离差系数是否与对应的均值系数相同来实现的。从表 4.8 可以看到,尽管均值系数没有一个统计显著,但从总体上而言,它们与离差系数都大不相同。针对模型差异的卡方检验只能为选择固定效应模型提供勉强的支持。对于随机效应模型,Wald 卡方值为 12.16,自由度为 6($p = 0.06$)。对于 GEE 模型,Wald 卡方值为 12.87,自由度为 6($p = 0.04$)。

总结

针对计数数据的固定效应模型,可以在因变量服从泊松分布或负二项分布的假定下进行估计。当每个个体只有两期观察时,固定效应泊松模型的条件最大似然估计,可以通过将泊松模型转变为解释变量为差分值的分组数据 logistic 回归模型来实现。当每个个体有两个以上的观察记录时,泊松模型的条件最大似然估计可以通过 Stata 中的 xtpoisson 命令完成。

无条件最大似然估计可以用标准的泊松回归软件完成,用虚拟变量来代表固定效应即可。与 logistic 回归不同,固定效应泊松模型的条件和无条件估计会产生相同的系数和标准误。不幸的是,标准误常常因为过离散问题而严重有偏。在 Stata 中,我用了自助标准误,以矫正过离散问题,其他一些软件包有计算上更简单的方法。

解决过离散问题更好的办法是估计一个带有过离散参数的负二项回归模型。不过,这种模型不能用条件最大似然法进行估计。无条件最大似然估计只要使用虚拟变量作为固定效应,在任何负二项回归软件中都能完成。

混合模型法在允许对时变自变量的固定效应系数进行估计的同时,也能对非时变自变量的效应进行估计。正如我们在第 2 章和第 3 章已经看到的那样,每个时变变量被分解成为两部分:一个分个体的均值(个体内均值)和一个相对于该均值的离差。回归模型包括这两套变量,同时还有非时变自变量。个体内的相依问题可以通过 GEE 估计或随机效应模型的最大似然估计进行处理。

第 5 章 | 事件史数据的固定效应模型

事件史分析这一名字指称的是一套被设计用来描述、解释或预测事件发生的统计学方法。在社会科学之外，这些方法通常被称作生存分析，主要是因为它们首先被生物统计学家发展出来，以分析死亡事件的发生。碰巧这些方法非常适合用来分析大量的社会现象，如出生、结婚、离婚、失业、晋升、被捕、迁移和反抗等。事件史分析还有很多其他名称，包括失败时间分析（failure time analysis）、风险分析（hazard analysis）、转换分析（transition analysis）和持续期分析（duration analysis）。

一般而言，一个事件可以被定义为发生在特定时刻的性质（qualitative）改变。若要应用事件史方法，就需要有事件史数据，也就是一种记录着事件何时发生在某个体或某些样本个体身上的纵贯记录。例如，让一群被抽作样本的妇女汇报她们所生的全部小孩的出生日期，你就可以得到一套可以用来分析出生事件（occurrence of births）的事件史数据。当然，如果你想进行因果分析或预测分析，你还得测量一些可能的解释变量，例如妇女自身的出生日期、教育水平、家庭收入、婚姻状态等等。

下面我们将这个例子变得更加具体。在 1995 年的全国家庭成长调查（National Survey of Family Growth, NSFG）中，一个能够代表全美国妇女的样本被要求报告她们曾经生育的所有小孩的出生信息（www. cdc. gov/nchs/nsfg. htm）。这里使用的是该数据的一个子样本，包括 6911 个至少生育过一胎的妇女。这些妇女总共报告了 14932 次活产事件。对每个生育事件，我都计算了出生间隔（birth interval），标为 DUR：从目前这次生育到下一次生育的时间长度（以月计算），如果没有后续生育被观察到，则为到调查当日的时间长度。这些出生间隔的潜在解释变量包括刻画当前生育特征的几个变量：

PREGORDR 生育次序(即胎次)(1到15)

MARRIED 如果生育时已婚则为1,否则为0

AGE 出生时母亲的年龄(以年计算)

PASST 生产费用全部或部分由政府援助基金(public assistance funds)支付则1,否则为0

NOBREAST 如果母亲不用母乳喂养小孩则为1,否则为0

LBW 如果所生为低体重儿则为1,否则为0

CAWSAR 如果生产为剖腹产(Caesarian section)则为1,否则为0

MULTIPLE 如果是多胞胎则为1,否则为0

另外还有一个变量COLLEGE,如果妇女受过一些大学教育(在调查时)则等于1,没有受过大学教育则等于0;再有一个变量BIRTH,如果出生间隔是以另一次生育事件结束的则为1,如果是由该调查作为结束的则为0——一个删失间隔。这一数据集有6911个删失间隔。每个妇女都有一个删失间隔,因为每个人的最后一个间隔都是被调查访问结束的。最后,变量CASEID是一个ID变量,其取值在同一个妇女的所有生育间隔记录中都是相同的。我们的目标是对生育间隔估计一个回归模型。

COX回归

分析事件史数据最流行的方法是Cox回归,这是以发展了比例风险模型(proportional hazards model)及估计这一模型的偏似然方法(partial likelihood method)的David Cox来命名的。在进行固定效应分析之前,我简单回顾一下这种方法。

Cox回归不是直接对间隔长度进行建模,其因变量是事件发生的风险(hazard)或瞬间的可能性。对于可重复发生事件,风险的定义如下:令$N_i(t)$为个体i在时间t以前发生的事件数,那么个体i在时间t的风险可以这样给定:

$$h_i(t) = \lim_{\Delta t \to 0} \frac{\Pr[N_i(t+\Delta t) - N_i(t) = 1]}{\Delta t} \qquad [5.1]$$

用文字来说，这个等式表达的是我们先考虑在某个极短的时间间隔 Δt 内一个新事件发生的概率。再构建这一概率对 Δt 的比率，然后取 Δt 趋近于 0 时这一比率的极限。对于重复发生事件，这一风险函数又被称为强度函数(intensity function)。

接下来，我们将这一风险建模成为解释变量的函数。令 $h_{ik}(t)$表示个体 i 发生第 k 次事件的风险，那么比例风险模型是如此给定的：

$$\log h_{ik}(t) = \mu[t - t_{i(k-1)}] + \beta x_{ik} \qquad [5.2]$$

其中x_{ik}是能够在不同个体间及事件间发生变动的自变量列向量，β 是系数行向量，$t_{i(k-1)}$是第($k-1$)次事件发生的时间，$\mu(\cdot)$是最近一次事件发生以来的时间长度的未定函数(unspecified function)。在这一模型中，我们假定 $\mu(\cdot)$对样本中的每一个个体都是同一个函数。

偏似然估计的一个显著特征，是它可以在不对函数 μ 做任何假定的情况下对β 进行估计。至于它是如何实现这一点的，可以参见笔者的另一本书(Allison, 1995)。在 Stata 中，Cox 回归是通过命令 stcox 完成的。表 5.1(头两列数

表 5.1 常规模型的 Cox 回归估计[a]

	系 数	常规标准误	稳健标准误	风险比率 (Hazard Ratio)
PREGORDR	−0.163	0.011	0.016	0.849
AGE	−0.065	0.003	0.003	0.937
MARRIED	0.221	0.029	0.030	1.247
PASST	0.137	0.029	0.029	1.147
NOBREAST	−0.270	0.023	0.023	0.763
LBW	−0.003	0.042	0.043	0.997
CAESAR	−0.116	0.030	0.028	0.890
MULTIPLE	−0.702	0.143	0.144	0.495
COLLEGE	−0.207	0.026	0.026	0.813

注[a] 除 LBW 的 p 值大于 0.90 外，其他系数的 p 值都小于 0.01。

字)给出了对上述出生间隔数据拟合 Cox 模型所得到的结果,这里将所有的出生间隔都当做相互独立的观察,也就是说,将每一个出生间隔都看作是来自总体中的不同妇女。除低出生体重外,所有变量对下一次生育的风险都有高度显著的影响。已婚或受政府援助的妇女有更高的生育风险。其他变量的系数都是负的。

要想得到有关这些结果的更具体解释,查看最后一列(标有“风险比”[hazard ratio])会有很大的帮助,它给出了系数估计值的指数幂。风险比解释起来几乎和 logistic 回归里的发生比率完全一样。例如 MARRIED 的风险比为 1.25。这意味着一个生育时已结婚的妇女再次生育的风险比没结婚的大 25%(在控制了模型中的其他变量后)。MULTIPLE 的风险比是 0.495,这意味着如果一个妇女生的是双胞胎,那其再次生育的风险将减半。对于 AGE 来说,风险比是 0.937,这表示(作为)母亲,其年龄每增加一岁,将使下一次生育的风险减小 $100*(1-0.937)=6.3\%$。

不过,上述结论有潜在的问题。有 69%的妇女每人至少为这一数据集提供了两个生育间隔,因此怀疑同一个人的多条观察之间存在一定相关是合理的。具体而言,很自然地就能想到,可能某些妇女的生育间隔就一直都比较短,而另一些的生育间隔一直都比较长。不考虑这种相依性将严重低估标准误和 p 值。

幸运的是,使用在前面章节中用过的稳健方差估计法修正标准误很容易(Therneau & Grambsch, 2000)。通过 vce(cluster caseid)获得的稳健标准误呈现在表 5.1 的第 3 列中。此处绝大部分的修正都很小,只有 PREGORDR① 的修正标准误例外,它比未修正时要大 37%。这样产生的修正 z 统计量只有未修正时的一半,不过仍然高度显著。

带固定效应的 Cox 回归

现在我们已做好准备将固定效应加到 Cox 回归模型中。与往常一样,这将允许我们控制所有稳定的预测变量,并处理好重复观察之间的相依问题。与此

① 此处原书为 PREGORDER,但根据上下文及输出表格可知应该为 PREGORDR。——译者注

前的几个固定效应模型一样，α_i 代表所有稳定的自变量的综合作用。我们的固定效应回归模型的第一形式是：

$$\log h_{ik}(t) = \mu(t - t_{i(k-1)}) + \beta x_{ik} + \alpha_i \qquad [5.3]$$

针对我们的生育间隔数据，方程 5.3 如何才能被估计出来呢？一种想当然的做法是，将标识每个妇女（除其中一个外）的虚拟变量（组）放到模型里去。这种方法在线性模型、泊松模型以及负二项模型中很奏效，但在这里会遇到严重的困难。首先，估计一个带有 6910 个虚拟变量的 Cox 回归在实际操作上就是个问题。[21]

更为根本性的麻烦在于估计如此众多的"伴随性参数"导致的可能偏差。在前面几章中，我们发现这种偏差在 logistic 回归模型中可能非常严重，但在泊松或负二项回归模型中并不如此。在其他地方（Allison，2002），我已经指出：Cox 模型在这一点上与 logistic 模型很像。当每个人的平均间隔数少于 3 个时，使用虚拟变量法估计固定效应模型所产生的回归系数偏差（偏离 0）约为 30%—90%，偏差大小取决于删失的水平（删失案例所占比例越高，产生的膨胀越大）。

幸好，另外有一种方法实行起来简单而且非常有效。像在 logistic 回归和泊松回归中使用的条件似然法，因为虚拟变量的系数并没有真正被估计出来而是被从似然方程中消除了。首先，我们调整方程 5.3，通过定义我们得到

$$\mu_i[t - t_{i(k-1)}] = \mu[t - t_{i(k-1)}] + \alpha_i$$
$$\log h_{ik}(t) = \mu_i[t - t_{i(k-1)}] + \beta x_{ik} \qquad [5.4]$$

在这一方程中，固定效应 α_i 被整合进了时间的未定函数中，这一函数现在被允许在每个个体上都不一样。注意，方程 5.4 与方程 5.2 中的常规 Cox 模型的唯一不同之处就在于 μ 的下标 i。这样，每个个体都有其自身的风险函数，这比只允许每个个体有自己的截距来得更加宽松。

方程 5.4 可以通过标准的 Cox 回归程序和广泛可得的分层选项（option of stratification）进行估计。分层（模型）允许不同的子群体有不同的基准风险函数（baseline hazard function），但同时限定系数在各个子群体中相同。它是通过为每个子群体建立一个偏似然函数，将所有这些似然函数连乘起来，然后在考虑

系数向量 β[1]的情况下，将结果似然函数最大化来完成的。利用 Stata 中的 stcox 命令，分层可以通过设置 strata(caseid)选项实现，这将意味着 6911 个妇女每个人都被当做一个独立的层。看起来层的数量似乎非常大，但 stcox 能够轻松搞定。

表 5.2 中模型 1 的结果，呈现出与表 5.1 不同的地方，这些差别值得注意。第一，COLLEGE 没有任何报告信息。像大多数固定效应方法(除了混合模型法)一样，我们不能估计那些在个体内不存在变化的变量的系数。从 COLLEGE 往上，我们看到多胞胎的系数与前面的估计差不多一样。但变量 CAESAR 的系数有些变小，并且统计上不再显著。低出生体重在前面是高度不显著，但这里 p 值小于 0.01。LBW 的风险比告诉我们，所生为低出生体重儿将使下一次生育的风险下降了 21%。母乳喂养的作用不管在大小还是在显著性上都有所下降。政府支持在前面高度显著，但在这里一点也不显著。婚姻状态的影响在这里差不多一样。年龄在统计上不再显著。而怀孕胎次的影响比前面要大得多，无论在大小还是在统计显著性上都如此。每多生一胎将会使生育下一胎的风险下降 50%。

表 5.2　固定效应模型的 Cox 回归估计

	模型 1			模型 2	
	系　数	标准误	风险比率	系　数	标准误
PREGORDR	−0.711**	0.034	0.491	−0.712**	0.034
AGE	0.007	0.011	1.007	0.007	0.011
MARRIED	0.181**	0.070	1.199	0.182**	0.070
PASST	0.077	0.069	1.080	0.076	0.069
NOBREAST	−0.128*	0.060	0.879	0.043	0.100
LBW	0.237**	0.081	0.789	−0.243**	0.081
CAESAR	−0.079	0.093	0.923	−0.080	0.093
MULTIPLE	−0.607**	0.218	0.545	0.590**	0.219
COLLEGE	(被剔除)			(被剔除)	
COLLBREAST				−0.267*	0.125

注：* $0.01 < p < 0.05$，** $p < 0.01$。

① 即保证 β 在各子群体中相同。——译者注

为什么固定效应估计会如此不同于常规 Cox 回归估计的结果呢？与所有固定效应模型一样，这一模型控制了所有稳定的自变量，因此早先表 5.1 中的有些结果可能是虚假的。如果非得让我在表 5.1 的常规结果与表 5.2 的固定效应结果之间做出选择的话，我会断然地选择后者。我们脑子里必须记住的是，在此处的分析中，每个妇女不同的出生间隔是与其自身进行比较。对于每一个妇女，我们的疑问是：为什么她的出生间隔中有一些会比另一些长或者短。例如，是因为她在某些出生间隔中处于结婚状态，而在另外一些未处于结婚状态吗？这一方法得出的答案，与考察为何有的妇女比另一些妇女倾向于有更长的出生间隔所得的答案是完全不同的。

固定效应模型的这一方面与 PREGORDR 变量尤其相关。在常规 Cox 回归中，这一变量对风险似乎有虚假的正向作用。在一个固定的时间区间内，生育次数多的妇女其生育间隔必然小。但通过固定效应分析，我们能够移除这一人为因素，这就是使负系数比原来大很多的原因。

和线性模型及 logistic 模型一样，尽管固定效应模型不能估计非时变变量如 COLLEGE 的作用，但它能够估计非时变变量与其他变量之间的交互作用。例如，我们可以估计一个含有 COLLEGE 与 NOBREAST 交互项的模型。这只需纳入 COLLEGE 和 NOBREAST 的乘积项作为预测变量之一就可以了。相应的结果在表 5.2 的模型 2 中。可以看到，这一交互项在 0.03 水平上统计显著。但如何对其进行解释呢？NOBREAST 的"主效应"代表当 COLLEGE $=0$ 时，也就说在未接受大学教育的妇女中该变量的作用。这个系数是正的但高度不显著。而 NOBREAT 在受过大学教育的妇女中的作用等于上述主效应加上交互项 $(-0.2659+0.0421=-0.22)$。使用 test 命令，我们可以发现两者之和显著不等于 0。因此，结论是在受过大学教育的妇女中，母乳喂养会增加随后一次生育的风险，但在其他妇女中这种影响不存在。

Stata 也能估计随机效应 Cox 模型，这一模型同样可用方程 5.3 设置，但假定 α_i 服从伽马分布且与 x_i 独立。这种类型的模型通常被称为"共享脆弱性 (shared frailty)"模型，其中 α_i（或者其指数幂形式）被描述为脆弱成分（frailty term）。其意思是说，有些个体比其他个体更加脆弱，因而更有可能经历该事

件。Stcox 命令中用于估计此种模型的选项是 shared(caseid)。不过,Stata 在试图对本出生间隔例子估计这一模型时遭遇了计算上的失败(computational failure),这显然是样本规模(过大)造成的。

但随机效应 Gompertz 模型的运算取得了成功(用 streg 命令完成),这一模型是方程 5.3 的特殊形式,方程中的 $\mu(\cdot)$ 被设定为一个线性函数。模型结果呈现在表 5.3 中。各项结果与表 5.1 中常规 Cox 模型的估计非常相似,即使没有进行稳健标准误修正。在 α_i 的方差估计值不显著地区别于 0 的情况下,这并不奇怪。当然,表 5.2 中的固定效应估计显著地不同于随机效应估计,再一次证明控制住未被观测到的异质性非常重要,即使在随机效应模型未能提供此种异质性存在的任何证据时也是如此。

表 5.3 随机效应 Gompertz 模型估计结果[a]

	系 数	常规标准误	稳健标准误	风险比率
PREGORDR	−0.163	0.011	0.016	0.849
AGE	−0.065	0.003	0.003	0.937
MARRIED	0.221	0.029	0.030	1.247
PASST	0.137	0.029	0.029	1.147
NOBREAST	−0.270	0.023	0.023	0.763
LBW	−0.003	0.042	0.043	0.997
CAESAR	−0.116	0.030	0.028	0.890
MULTIPLE	−0.702	0.143	0.144	0.495
COLLEGE	−0.207	0.026	0.026	0.813

注[a] 除 LBW 的 p 值大于 0.90 外,其他所有系数的 p 值都小于 0.01。

附加说明

尽管固定效应 Cox 模型有上述吸引力,但它同样有常见的缺陷。和其他固定效应方法一样,与常规分析相比,固定效应 Cox 模型的统计力会大大降低。在本例中,只有一个生育间隔的妇女都被排除在分析之外,因为这些间隔无法与其他间隔进行比较。这排除了 2109 个生育间隔。第二,在只有两个生育间隔

的妇女中，如果第二个生育间隔（总是有删失的）小于第一个，那两个生育间隔都会被排除在分析之外。原因如下。假定第一个生育间隔是 28 个月，而第二个间隔是 20 个月。在建构发生于第 28 个月的生育事件的偏似然函数时，计算公式将在相同的时间点寻找其他“处于风险中”的间隔（来自同一个妇女）（looks for other intervals that are “at risk” of the event at that same time）。但另一个生育间隔在第 20 个月时删失了，这样，对于那个生育间隔，该妇女已经不再处于在第 28 个月发生可观察生育事件的风险中了（the woman is no longer at risk of an observable birth at 28 months）。因此，前述出生没有对象可进行比较，该妇女也就被排除在偏似然函数之外。在 NSFG 数据中，此种间隔的排除又导致 1468 个案例被损失掉。

最后，即使对那些保留下来的观察记录，固定效应方法也根本就没有考虑不同妇女间的变异信息，而只使用了妇女内变异。因此，如果某个协变量在不同妇女之间存在很大差异，而每个妇女历时变化很小时，那么该变量的系数将不能被可靠地估计出来。例如，变量 PASST 的 80%的变异在于不同妇女之间，属于妇女内的变异只有 20%。因此毫不奇怪，表 5.2 中其系数的标准误，与表 5.1 相比，是后者的两倍多，因为后者的标准误是基于妇女间及妇女内两种变异计算出来的。

除了常见的固定效应模型的不足之外，固定效应 Cox 回归还容易受特定类型的变量的影响。这些问题最有可能在出生间隔研究中出现的这类数据结构下发生。在这种数据结构下，每个个体被观察了一段固定的时间，在这一段时间内，可能有多次事件发生，但只有最后一个间隔是删失的。张伯伦（Chamberlain, 1985）认为这种结构违反了似然估计的基本条件，因为一个间隔被删失的可能性取决于前一个间隔的长度。

在一个仿真研究中（Allison, 1996），我已经指出这种违背对于绝大多数自变量不会产生严重的问题，但在估计刻画以往事件特征的变量的系数时会导致偏差。具体地讲，固定效应偏似然估计倾向于让以往事件的数量、以往间隔的长度对风险产生负的影响，即使这些变量并没有真正的影响。这无疑与表 5.2 中的结果相一致，表中胎次对下一次生育的风险有很强的副作用。这一问题在每个个体的平均事件数少，删失间隔在所有间隔中所占比例高的情况下最严

重。不过，前面我已经指出，对此前事件数的影响的估计，常规 Cox 回归的偏差可能更大，只是偏向另外一个方向。

Cox 回归混合模型法

在前面的章节中，我们看到可以通过将时变自变量分解成为具体单位的均值以及相对这些均值的离差，然后将所有这些变量纳入常规回归分析，可能的话，再修正同一个体的多次观察之间的相依性就能复制或近似地模拟固定效应分析的结果。不过，由于某种并不太清楚的原因，这一方法看起来在 Cox 回归中不大好使。例如，如果我们将混合法用于这一生育间隔数据，有几个变量的系数及 p 值与表 5.2 中的存在天壤之别。我对 Cox 回归混合法的仿真研究也很让人沮丧。因此，我无法为事件史分析推荐、介绍混合模型法。

非重复性事件的固定效应事件史法

固定效应 Cox 回归要求样本中至少有一些个体经历一个以上的事件，这样个体内的比较才成为可能。显然，这种方法无法应用于不可重复的事件，例如死亡。不过，在某些条件下，通过应用条件 logistic 回归，并把时间看作离散的，可能可以对非重复性事件进行固定效应分析。在流行病学文献中，这类分析叫作病例交叉研究(case-crossover study)(Maclure, 1991)，尽管我这里描述的实现方式与流行病学通常所做的存在一些差异。

与通常一样，我将从一个经验的例子开始。设想我们要回答下面的问题：妻子的去世是否会增加丈夫去世的风险？这是一个很难有信心回答的难题，因为丈夫的去世与妻子的去世之间的任何相关都可能是共同的环境特征影响下的结果。他们中的大多数都已经在相同的住所(house)、相同的邻里环境(neighborhood)中共同生活了很长一段时期。而且，他们倾向于来自相同的社会经济背景，有着类似的生活方式(lifestyles)。除非我们能够控制这些共同点(commonalities)，否则任何观察到的一个配偶的死亡与其另外一个的相关都有

可能是虚假的。因此，非常需要将固定效应分析作为一种方法，来控制所有稳定的、未被测量的解释变量。

为了回答这一问题，我将分析一个含有 49990 对已婚夫妇的数据，数据中夫妇双方在 1993 年 1 月 1 日[22]时都还健在，并且都至少已有 68 岁。截至 1994 年 5 月 30 日期间，已亡夫妇的死亡日期都是已知的。在这 17 个月期间，丈夫中有 5769 人死亡，而妻子中有 1918 人过世。我们将时间看作是由离散的单位组成的，在这里是以日计算，列举出来就是 $t=1, 2, 3$ …… 令 p_{it} 表示在前一天仍然活着的条件下丈夫 i 在 t 日死亡的概率，再令 $W_{it}=1$，如果妻子 i 在 t 日还活着的话，否则为 0。

我们将用一个 logistic 回归模型来表示妻子的存活状况对其丈夫死亡的概率的影响：

$$\log\left(\frac{p_{it}}{1-p_{it}}\right)=\alpha_i+\gamma t+\beta W_{it} \qquad [5.5]$$

其中 γt 表示时间对死亡的对数发生比的线性作用，α_i 表示所有未被测量的在各个时期保持稳定的变量的固定效应。注意：模型中没有放入非时变预测变量，因为它们的作用已经被整合进了 α_i 项中。

现在我们试图用第三章中描述过的条件最大似然法对这一模型进行估计，这种方法将所有的 α_i 都从估计方程中消掉。下面是这一数据集的构成方式。对于去世了的男人，夫妇们被观察的每一天都创建了一条单独的观察记录，从第一天(1993 年 1 月 1 日)到去世那天为止。对于这些夫妻-日记录(couple-days)，因变量 Y_{it} 被编码为 0，如果该男性在当天还健在的话；如果在那一天他去世了，则编码为 1。这样，一个在 1993 年 6 月 1 日去世的男性将提供 152 个夫妻-日；其中 151 个 Y_{it} 的取值为 0，最后一个的取值为 1。解释变量 W_{it} 在妻子活着的日子里都被编码为 0，妻子已经过世了的日子里都被编码为 1。尚未去世的男性没有创建记录，因为在二分结果变量的固定效应分析中，没有发生变化的个体对似然函数没有任何贡献。本工作数据中夫妻-日的总数为 1377282。和第 3 章中描述的一样，模型可以通过 Stata 的 xtlogit 或 clogit 命令进行估计。

不幸的是，对于这两个命令，用来使似然函数最大化的计算公式都不能收

敛。对数似然值很快就变成 0，且迭代序列延续不断、没个尽头。收敛失败的原因在于每对夫妇那串观察记录的因变量都是由一连串 0 跟上一个 1 组成的。也就是说，事件总是发生在最后一个观察单元。因此，时间或时间的任何单调递增函数（例如时间的对数，或时间的平方根）都将完美地预测该夫妇的结果，从而无法得到该协变量或模型中任何其他协变量的最大似然估计。在 logistic 回归文献中，这一问题被叫作完全分离（complete separation）（Albert & Anderson，1984；Allison，2004）。[23]

事实上，对于我们这个死亡事件的例子，不收敛问题并不局限于时间变量的原因。即使把时间移出模型，我们得到的仍是不收敛（尽管现在的问题不是完全分离而是半完全分离[quasi-complete separation]）。因为 W_{it}，即妻子是否死亡这一虚拟变量，随着时间推延而增加，但从不减小，它完美地预测了最后一天的死亡事件的发生。因此，它的系数在计算公式每迭代一次时都会变得更大。

克服这个问题的一种方法是把 W_{it} 改进成为一个标识，指示妻子是否还在世，比方说，过去 60 天内过世。当妻子去世时，这个协变量从 0 变化为 1，但在第 60 天之后又变回成为 0（如果这个丈夫还活着的话）。通过估计时间窗口大小不同的模型（model with varying windows of time）可以提供有用的信息，描述妻子死亡的影响如何开始、到达顶峰以及结束。

表 5.4 的上半部分同时给出了使用几个不同时间窗口的情况下（但没有包括时间本身的影响），衡量妻子的过世对于丈夫去世的影响的发生比率的固定效应估计。在所有情况下，发生比率都超过 1.0，并且 60 天间隔和 30 天间隔下都统计显著。对于后一种情况，在妻子死后 30 天内，丈夫死亡的发生比是其他时间发生比的 2 倍①。表 5.4 的下半部分给出了常规 logistic 回归得到的发生比率，没有控制稳定不变的、未被观察的协变量。与上半部分相比，这一部分的发生比率都要小，而 p 值都要高。

---

① 这里实际上是将妻子死后 30 天内丈夫死亡的发生比与其他时间（包括妻子死前或没死，以及妻子死了 30 天之后 3 种情况）丈夫死亡的发生比进行比较。——译者注

表 5.4 在不同时段内以妻子的死亡事件对丈夫的死亡进行预测的发生比率

		妻子死亡多少天以内				
		15 天	30 天	60 天	90 天	120 天
固定效应估计	风险比率	1.26	1.96	1.61	1.27	1.26
	p 值	0.54	0.006	0.03	0.24	0.25
常规估计	风险比率	1.13	1.56	1.21	0.97	0.93
	p 值	0.71	0.04	0.29	0.87	0.61

尽管这些结论非常有趣，但危险在于，模型没有对历时的变化进行控制。这不仅仅是一个技术问题，还是一个能够对从病例交叉研究中得出的任何结论都构成严重危害的问题(Creenland, 1996; Suissa, 1995)。对于我们的例子来说，只要妻子的死亡发生率有随着观察期延长而增加的趋势，就有可能导致丈夫的死亡与妻子的死亡(不管如何编码)之间的虚假相关。直观地讲，原因在于：丈夫的死亡总是出现在每对夫妇的观察序列的结尾，因此，任何倾向于随着时间而增加的变量，看起来都会增加丈夫死亡的概率。

现在我们来考虑另外一种固定效应模型，它看起来能够解决因未控制时间的影响而造成的问题。休莎(Suissa)发明了一种方法，并把它叫作"案例—时间—控制"设计，这种方法的关键创新之处，在于将条件 logit 模型中的因变量与自变量进行调换的计算策略(computational device)。这使得在模型中纳入对于时间的控制成为可能，而这种控制在病例交叉法中是无法实现的。

众所周知，当因变量和自变量都是二分变量时，发生比率是对称的——调换因变量和自变量将得到相同的结果，即使模型中还有其他自变量。[24]在案例—时间—控制法中，工作因变量(the working dependent variable)是二分协变量——在我们这个例子中，是妻子是否在过去一段日子里死亡。自变量是标识事件(丈夫的死亡)是否在给定日期发生的虚拟变量，以及时间的某种适当形式，如一个线性函数。同样，估计的还是一个条件 logistic 回归，且将每对夫妇都单独作为一层对待。在这一方程中，将时间作为协变量纳入不存在问题，因为工作因变量不是时间的单调函数。

在休莎方法的方程式中，有必要纳入所有个体的数据，包括经历了事件的人以及被删失了的人。不过，他的模型只是针对每个个体仅有两个时间点的数据发展出来的，一个事件时段、一个删失时段。在那种情况下，如果样本仅仅局限于经历了事件的人，那协变效应和时间效应将完全混合在一起。而删失案例则能够提供协变量受制于时间的有关信息，这些信息与事件的发生不存在混淆。

不过，我们的数据集(有可能很多其他数据也)在不同时点上对每个个体有多个"控制"。这消除了时间与事件发生(丈夫的死亡)之间的完全混合，使得我们可以将案例—时间—控制法只用于未被删失的案例。在很难或无法收集到未经历事件的人们的信息时，这是一个极大的好处。如果估计时未包括删失案例，这个模型的唯一限制是，我们无法估计一个时间影响完全随意的模型，也就是说，模型中不能带有标识每个时点的虚拟变量。

当然，如果删失案例的信息是可得的(如在我们这个数据集中)，那么纳入它们，可以得到更准确的对时间的影响的估计。不过，即使删失案例是可得的，将分析限制于事件经历者仍具有潜在的优势。案例—时间—控制法假定协变量对于时间的依赖(即时间对于协变量的影响)在经历了及未经历事件的人身上是一样的，这一点为人所批评(Greenland, 1996)。如果分析的数据仅限于事件的经历者，那此种批评也就毫无力道了。

对于死亡事件数据，工作数据集与前面的一样，从开始到丈夫死亡或者删失当天，每个观察者每天都有一条记录。因为条件 logistic 回归要求每个条件层(conditioning stratum)的因变量都有所变动，我们可以将妻子没有在丈夫之前死亡的夫妻案例删除，而不存在信息损失。

本工作数据集中有 39942 个夫妻-日，仅来自于 126 对夫妇。这是丈夫死亡且妻子先于丈夫死亡的夫妻数。尽管这只是原样本 49990 对夫妻中很小的一部分，但使用固定效应方法时，只有这些人含有妻子的死亡对丈夫死亡的影响的信息。这是一个问题么？如果同一个模型(带着同样的系数)适用于总体中的每一个人，那就不是一个问题。但如果不同子群体的模型不一样，那从这 126 对夫妻获得的结果就只能准确地描述他们而已，但不能描述整个总体。

工作模型的定义如下。令 H_{it} 表示丈夫 i 在 t 天是否死亡的虚拟变量，而 P_{it} 表示在 t 天之前的一定时日内妻子死亡的概率。Logistic 回归模型为：

$$\log\left(\frac{P_{it}}{1-P_{it}}\right)=\alpha_i+\beta_1 H_{it}+\beta_2 t+\beta_3 t^2 \qquad [5.6]$$

表 5.5 以妻子之死对丈夫的死亡进行预测所得不同时段内的风险比率（案例—时间—控制法）

	15 天	30 天	60 天	90 天	120 天
风险比率	1.26	2.08	1.74	1.28	1.11
p 值	0.54	<0.004	0.01	0.25	0.63

尽管也可以使用其他函数，但这一模型允许纳入时间的二次项作用（quadratic dependence on time）。

表 5.5 给出了不同时间窗口下发生比率的估计。结果与表 5.4 中的非常相像，后者使用的是病例-交叉法。证据再次表明妻子的死亡对于丈夫死亡的风险的影响受到时间的限制，妻子死亡两个月后其影响将大大降低。

尽管我们的工作因变量是妻子的死亡，但发生比率得解释成妻子的死亡对于丈夫死亡的发生比的影响。这是由于观察的时间顺序的原因——妻子的死亡总是发生在丈夫之前。如果我们的目标是估计丈夫的死亡对于妻子死亡的影响，那我们得构建一个完全不同的数据集，包含妻子死前的夫妻-日记录，而不是妻子死后的。

在这个例子里，我们只估计了一个二分协变量（妻子在一定的时日内的死亡）对于不可重复事件（丈夫的死亡）的影响。这种方法允许我们控制所有固定变量。但是，假设我们想控制时变自变量，如吸烟状况。仿真研究（Allison & Christakis，2006）发现，附加解释因素作为自变量，可以直接纳入等式 5.6 所设定的 logistic 回归模型。尽管附加自变量的系数不是这些变量对丈夫死亡的影响的无偏估计，但纳入这些解释因素后，能够得到妻子的死亡对于丈夫的死亡的影响（方程 5.6 中的 β）的近似无偏估计。假如我们要估计吸烟状况对丈夫的死亡的影响，那我们就得使吸烟的概率成为方程 5.6 中的因变量，另外有可能再

纳入妻子的存活状态作为协变量。即使吸烟状态有两个以上类别，这种程序仍然有效，只是这时等式 5.6 需要被设定成为一个多分类 logistic 回归。不过，我听说没有办法把案例-时间-控制法推广到对定量自变量的影响进行估计。

总结

事件史数据的固定效应回归分析通常要求每个个体有多个的、重复的事件。与我们在 logistic 回归中看到的一样，使用虚拟变量法估计固定效应，常常会导致对于其他变量的系数的有偏估计。这一伴随性参数问题在使用 Cox 回归时可能被避免，这种方法利用分层法将固定效应从偏似然函数中消除，即使是在分层数目很大的情况下仍然具有计算效率。在大多数情况下，分层法都能产生近似无偏的估计。

与其他固定效应方法一样，分层 Cox 回归在统计功效上也会遭受巨大的损失。自然的，只有一条观察记录的个体不能给分析提供任何信息。即使一个个体有一条删失记录和一条非删失记录，只要删失记录的时间区间相对较短，这个个体的两条记录也会被剔除在分析之外。最后，只有个体内的变异信息被用来估计各个系数。由于一些目前我们还不太清楚的原因，混合模型法——它在线性、logistic 和计数数据回归中运行良好——在 Cox 回归中无法产生正确的结果。

尝试对非重复性事件进行固定效应回归分析会遇到严重的困难。基本策略是把时间看作是离散的，然后分别针对每个人的各个被观察的离散时间点创建一条独立的记录，从开始观察一直到事件发生或者删失时为止。对于每一条记录，都有一个二分因变量，如果事件在该时间点发生则编码为 1，否则编码为 0。最后一步是，对这一因变量估计一个条件 logistic 回归，并将每个个体单独作为一层，而自变量为在不同时点上有所变化的变量。这种具有吸引力的方法的一个根本问题是，如果时间（或者时间的任何单调函数）被作为解释变量，那模型会因为分离问题（separation）而得不到收敛。原因在于：事件总是发生在每个个体的观察序列的末尾，从而使得时间能够完全预测事件的发生。

尽管未纳入时间(变量)的模型确实能够被估计出来,但因为时间对于(事件)风险以及自变量的影响未被控制住可能是有偏的。一种解决途径是案例-时间-控制法,这种方法好像对估计分类协变量对于风险的影响很管用。这种方法的创新之处在于调换了条件 logistic 回归中因变量与自变量的角色,从而使得在模型中纳入作为协变量的时间成为可能。

第 6 章 | 固定效应结构方程模型

在第 2 章中，我们考虑了几种估计线性固定效应回归模型的不同方法。在这一章，我们将展示如何把固定效应回归当作带一个潜变量的线性结构方程模型来估计。为什么我们还需要另外一种方法来估计同一个模型呢？答案是：通过将模型置于一种结构方程框架，我们能够得到一些通过常规计算方法难以或者不能得到的结果。具体地讲，我们可以：

(1) 估计固定效应和随机效应的折中模型，

(2) 构建对固定效应与随机效应进行比较的似然比检验，

(3) 估计两个反应变量间存在相互作用的固定效应模型，

(4) 估计反应变量为时滞值的固定效应模型，

(5) 估计潜变量带有多个指标的模型。

我之所以在这里为这种方法单辟一章，是因为其数据结构及概念框架与第 2 章中大多数方法所用的非常不一样。我首先将解释如何用结构方程软件估计第 2 章中描述过的随机效应模型。然后，我们会考察如何对这种模型进行调整，以形成固定效应模型。

随机效应作为潜变量的模型

在第 2 章中，随机效应模型被设置为：

$$y_{it} = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i + \varepsilon_{it} \qquad [6.1]$$

其中 y_{it} 是个体 i 在时间 t 的反应变量取值，x_{it} 是时变自变量向量，z_i 为非时变自变量向量，α_i 表示随机效应，ε_{it} 为随机扰动项。我们假定 α_i 和 ε_{it} 代表独立正态分布变量，其均值为 0，方差恒定。我们还假定，至少是在现在，这些随机成分都与 x_{it} 及 z_i 相互独立。

众所周知(Muthén, 1994)，如方程 6.1 所示的随机效应模型可以用结构方程模型(SEM)来表示，后者可以用众多被设计用来估计此种模型的软件(如 LISREL, EQS, MX, Mplus，或者 Amos)之一进行估计。不幸的是，Stata 中没有估计这种模型的命令。[25]在这里，我是用 Mplus(www. statmodel. com)来估计本章讨论的模型的。从概念上讲，我们认为方程 6.1 给每个时间点都设定了单独的公式，但限定各个时点对应的回归系数相同。随机项 α 和 ε 被当成潜变量。不同时点有不同的 ε，但是各个时点的 α 却是相同的。

SEM 通常用通径图来表示(Kline, 2004)。图 6.1 是一个有三期数据、一个时变自变量的模型的通径图。在 SEM 的通径图中，按照惯例，通常把直接观察变量放在矩形框中，而把潜变量放在圆圈或者椭圆中。直线单向箭头表示一个变量对另一个变量的直接因果作用，而曲线双向箭头表示两个外生变量之间的二元相关(bivariate correlation)。(用结构方程模型的术语来说，内生变量是那些至少在一个方程中作为因变量的变量。外生变量是那些未在任何方程中作为因变量的变量。)

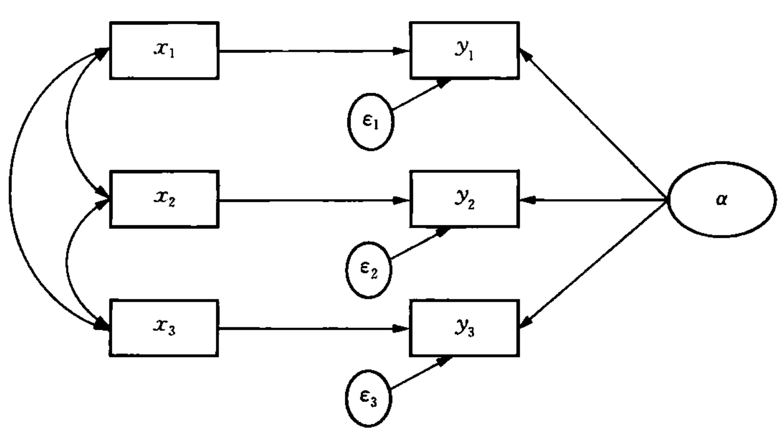


图 6.1 三期数据随机效应模型的通径图

在第 2 章中，我们使用 Stata 中的 xtreg 命令估计了公式 6.1 中的模型，当时是应用在 NLSY 数据上，该数据含有 581 名儿童在 3 个不同时期的观察记录。当时的工作数据包含了每个小孩的 3 条记录，总共有 1743 条记录。因变量是对反社会行为(ANTI)的测量。自变量包括两个时变变量：贫困状况(POV)和自信水平(SELF)，还有几个非时变变量。

要像结构方程模型一样估计该模型，我们将使用每个小孩只有 1 条记录、同一变量在 3 个时点的观察结果对应不同变量名的原始数据形式。使用 Mplus 实现这一估计的程序代码在附录 2 中。

在编写 SEM 程序来估计随机效应模型时有几点需要谨记：

(1) 很多 SEM 程序包估计模型时，默认情况下只使用协方差矩阵信息，这时你得不到截距(公式 6.1 中的 μ_1 表示)的估计结果。如果想得到截距，你需要通过恰当的设置把均值整合到分析当中。但是这并不会改变回归系数。

(2) 这个模型被设置成为三个独立的方程，分别针对 ANTI90，ANTI92，ANTI94。三个方程对应的系数被限定一致。在 Mplus 中，这是通过在自变量名后面的括号里放上数字来实现的，希望被限定相等的参数后面放相同的数字。放宽这些限定就相当于允许自变量与时间之间存在交互作用。

(3) 同时需要限定 ε_1，ε_2 和 ε_3 的误差方差(error variance)在三个方程中保持相等。

和大多数 SEM 程序一样，Mplus 也会产生大量的输出结果。这一输出至关重要的部分——回归系数、标准误以及统计检验情况——已经呈现在表 6.1 中。这些估计结果应该与表 2.5 中采用 xtreg 命令产生的结果进行比较。两套回归系数及标准误实质上是一模一样的。[26]

现在我们有了一种使用 SEM 软件估计随机效应模型的方法，其产生的结果与 Stata 中 xtreg 命令产生的相同。不过，这种方法有几个重要的不足。首

先，和 xtreg 命令不一样，这种方法很难用于非平衡数据。如果样本中每个个体的重复测量次数相等，则数据是平衡的。相反，如果我们的样本中有些小孩存在缺失值，比如说在变量 ANTI94 上，有些 SEM 软件要求把这样的个体记录全部删除。幸好，Mplus 以及其他大多数 SEM 程序包现在都有带缺失数据情况下的最大似然估计选项，从而能够处理此种非平衡数据。第二，尽管可能，但要想设置模型以处理时间的线性作用、与时间的线性交互作用、或者随机系数也会非常麻烦(Muthén & Curran, 1997)。相反，这在 xtreg 及大多数随机效应软件中很容易处理。

表 6.1 NLSY 数据的结构方程模型

	随机效应		固定效应		组合模型(Compromis)	
	系　数	标准误	系　数	标准误	系　数	标准误
SELF	−0.062**	0.009	−0.055**	0.011	−0.062**	0.009
POV	0.247**	0.080	0.112	0.093	0.111	0.093
BLACK	0.227	0.125	0.269*	0.126	0.269*	0.126
HISPANIC	−0.218	0.138	−0.198	0.138	−0.201	0.138
CHILDAGE	0.088	0.091	0.089	0.091	0.090	0.091
MARRIED	−0.049	0.126	−0.022	0.126	−0.025	0.126
GENDER	−0.483**	0.106	−0.476**	0.106	−0.479**	0.106
MOMAGE	−0.022	0.025	−0.026	0.025	−0.025	0.025
MOMWORK	0.261	0.114	0.296**	0.115	0.295**	0.115

注：* $0.01 < p < 0.05$，** $p < 0.01$。

但是，SEM 方法也有一些重要的优点。第一，它可以综合随机效应模型与带有多个指标的潜变量模型。这些潜变量既可以是自变量也可以是因变量。有关多指标潜变量模型的很好的入门介绍可以在克兰(Kline, 2004)或者哈彻(Hatcher, 1994)的著作中找到。第二，如我们将在下一节看到的，基于 SEM 框架的随机效应模型可以被扩展到用于估计固定效应模型，这种估计是通过允许随机效应模型与固定效应模型进行比较或折衷的方式来实现的。

固定效应作为潜变量的模型

如在第 2 章中已经提到的，基本的随机效应模型实际上只是固定效应模型的一种特殊形式（Mundlak，1978）。随机效应模型假定 α_i 与时变预测变量的向量x_{it}不存在相关。而固定效应模型允许 α_i 与x_{it}的元素存在任意的相关。图 6.2 展示了一个简化的只有一个时变自变量的固定效应模型。这个通径图与图 6.1 的唯一差别在于：α 和 x 之间多了表示相关的曲线箭头。

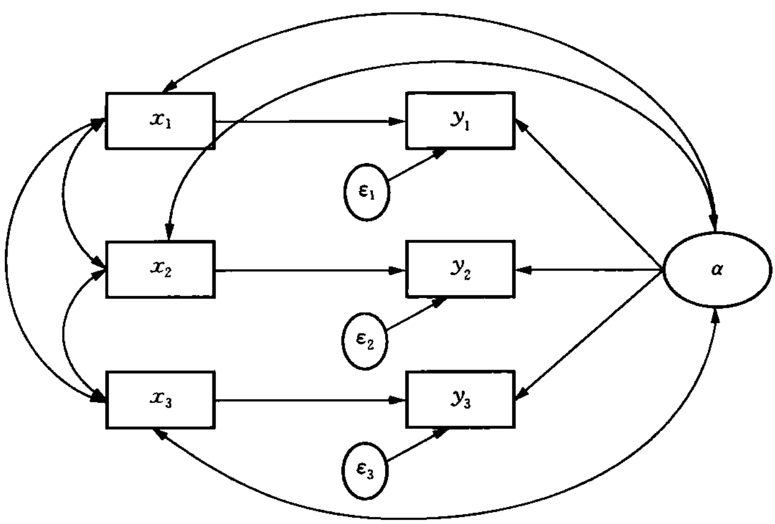


图 6.2 三期数据固定效应模型的通径图

这些新加的相关通过简单地设定潜变量与时变自变量之间的相关很容易就能合并到 SEM 软件中（Allison & Bollen，1997；Teachman，Duncan，Yeung & Levy，2001）。注意，潜变量不能与任何非时变自变量如 GENDER 或 MARRIED 存在相关。试图这样做将导致不能识别的模型（underidentified model），通常会产生错误或警告信息。

固定效应模型的系数估计及相关的统计量呈现在表 6.1 的中间两列中。我们首先来看 SELF 和 POV 的系数及标准误，可以看到它们与表 2.5 中使用 xtreg 命令的固定效应选项估计的结果一模一样[27]。它们与表 2.8 中采用混

合模型法所得结果也完全一样。

与混合模型法一样，表 6.1 也给出了非时变变量的系数估计。不过，表 6.1 中这些变量的系数估计值和统计检验结果与表 2.8 中的估计值及统计检验结果大不一样。例如，表 6.1 中 MOMWORK 的系数明显地统计显著，但在表 2.8 中同样明显地不显著。哪一个更好呢，混合模型估计还是 SEM 估计？这得看情况。仿真结果(这里没有呈现出来)强烈地表明，当非时变协变量 z 与未被观测的异质性因素 α 之间的相关为 0 时，SEM 产生的估计值近似无偏，而混合模型法产生的估计结果将可能大幅有偏(substantially biased)。相反，当 z 与 α 相关时，两种估计都会有偏，但 SEM 估计结果的偏差会比混合模型法的偏差更大。

既然我们已经同时有了固定效应和随机效应两种形式的结构方程模型，那么很容易就能产生一个对二者进行比较的似然比统计量(statistic)。对于每一个模型，输出结果都会包括一个卡方统计量和相应的自由度。这个统计量将模型的整体拟合水平与能够完美地重生(perfectly reproduce)所有变量的协方差矩阵的饱和模型进行比较。对于随机效应模型，卡方值为 84.42，自由度为 34。而对于固定效应模型，卡方值与对应的自由度分别为 66.45 和 28。两者之间的差异是一个取值为 17.97，带 6 个自由度的卡方量。这 6 个自由度对应的是固定效应模型下允许的另外 6 个相关。这一卡方量的 p 值为 0.006，表明我们应该拒绝随机效应模型而选择固定效应模型。这与我们在第 2 章中应用 Stata 产生的 Hausman 检验及检验均值变量系数与对中值系数是否相等的检验达成的结论相同。[28] 与混合模型法中的检验一样，这里计算的似然比检验拥有比 Hausman 检验更好的统计性质，比方说，后者在一些数据构造下可能取负值。

固定效应和随机效应的折衷

在上一节，我们是以随机效应模型作为开始，然后允许随机效应 α 与时变解释变量之间所有可能的相关的方式获得固定效应模型的。但可能并非所有那些相关都真实存在。表 6.2 给出了使用 Mplus 产生的 α 与时变变量之间的相

关系数与协方差估计值。看起来，α 与 SELF 变量的相关系数很小，而且在统计上不显著，但与 POV 变量的相关系数要大一些，且 3 个中有两个统计显著。这说明我们可以将 α 与 SELF 的相关系数设置为 0，而不会明显降低模型的拟合水平。这样做是令人满意的，因为它将使我们对 SELF 系数的估计同时基于个体内及个体间的变异，得到的标准误将更小些。[29]

表 6.2　α 与时变自变量之间的相关系数

	相　关	z 统计量
SELF90	−0.006	−0.77
SELF92	−0.0146	−1.71
SELF94	−0.008	−1.01
POV90	0.123	3.34
POV92	0.049	1.33
POV94	0.095	2.49

这在 Mplus 中很容易就能实现，产生的结果呈现在表 6.1 的右边两列。POV 的系数及 t 统计量与我们在固定效应模型中得到的差不多相同。另一方面，SELF 的系数比纯固定效应模型的略微大些，而其标准误要小 20%左右。取这两个模型的卡方之差，我们得到自由度为 3 的卡方值 3.00。这显然在统计上不显著，这表明我们不能因为喜欢较复杂的模型而拒绝较简单的模型（将 3 个相关系数设定等于 0 的那个）。

带滞后自变量的交互效应

至此，我们已经看到第 2 章中很多固定效应和随机效应模型同样也能用 SEM 软件进行估计，而且这种方法既有优点也有不足。下面我们来考虑一些远远超出第 2 章的重要固定效应模型，它们在结构方程框架下估计起来非常方便。这些模型违背了第 2 章的严格外生性假定，这一假定表述的是，在任意时点 t 及 t'，x_{it} 都在统计上独立于 $\varepsilon_{it'}$。这种情况的出现或是因为 x_{it} 受早先时点的 y 的影响，或是因为 x_{it} 的某个元素就是早先时点的 y 本身（时滞因变量）。这些模型非

常重要，因为它们为增强我们确定相互关联的变量间的因果方向的能力提供了可能。

假设我们观测了两个变量，x 和 y，已知两者存在相关。我们想知道究竟是 x 导致了 y 还是 y 导致了 x（或者同时存在）。两个变量都在几个时点上得到观察。假定有下面的模型：

$$y_{it} = \mu_t + \beta x_{i(t-1)} + \alpha_i + \varepsilon_{it}$$
$$x_{it} = \tau_t + \delta y_{i(t-1)} + \eta_i + \upsilon_{it} \qquad [6.2]$$

这一模型所说的是：y 受前一个时点的 x 的影响，而 x 受前一个时点的 y 的影响。模型还包括固定效应 α 和 η，它们代表任何及所有非时变解释变量对每个变量①的作用。我们同样可以纳入其他时滞型时变自变量和非时变变量，但这将使得我们的讨论变得不必要的复杂。

这一模型如何估计呢？如果观察恰好只有 3 个时点，那模型可以通过取一阶差分，然后分别对每个方程应用一般最小二乘法的方式进行估计：[30]

$$y_{i3} - y_{i2} = (\mu_3 - \mu_2) + \beta(x_{i2} - x_{i1}) + (\varepsilon_{i3} - \varepsilon_{i2})$$
$$x_{i3} - x_{i2} = (\tau_3 - \tau_2) + \delta(y_{i2} - y_{i1}) + (\upsilon_{i3} - \upsilon_{i2}) \qquad [6.3]$$

当超过 3 个时点时，第 2 章使用的方法（纳入标识每个个体的虚拟变量或者相对于均值的离差）看起来能够完成这一工作。不幸的是，由于存在往复效应，在固定效应估计中使用的差分值必然与回归方程中的误差项相关，从而导致有偏的估计（Woodridge，2002）。幸好，通过将固定效应并入结构方程模型，能够帮助我们规避这些难题。

当模型进一步扩展到允许纳入因变量（内生变量）的时滞值时，会出现更严重的困难：

$$y_{it} = \mu_t + \beta_1 x_{i(t-1)} + \beta_2 y_{i(t-1)} + \alpha_i + \varepsilon_{it}$$
$$x_{it} = \tau_t + \delta_1 x_{i(t-1)} + \delta_2 y_{i(t-1)} + \eta_i + \upsilon_{it} \qquad [6.4]$$

---

① 此处指 x 及 y。——译者注

如果我们将固定效应(α 和 η)排除在外,这一模型就是社会科学中有名的两期一双变量面板模型(two-wave, two-variable panel model)或交叉滞后面板模型(cross-lagged panel model)。

在计量经济学文献中,带滞后因变量的面板模型被称为动态模型。它们因对常规估计方法造成严重困难而出名,已经有几种可供选择的办法来应对该模型(Baltagi, 1995; Honoré, 1993; Honoré & Kyriazidou, 2000)。这些方法一般依赖于工具变量(IV)框架下对滞后变量的使用。其中最有名的是 Arellano 和 Bond(1991)提出的,通过在 Stata 中用 xtabond 命令来实现。不过,Lancaster(2000)将 IV 法描述成为"临时的",而且"只是因为计量经济学家不懂得如何正确使用似然法"才被使用的。

确实,动态固定效应模型的最大似然估计可以通过使用 SEM 软件直接估计。尽管这种方法的性质尚未得到分析性的深入研究,我所做的仿真研究(Allison, 2000)显示,这种方法能够很好的再生(recover)等式 6.4 所示模型的参数。

作为示例,我分析了美国 1983, 1989, 1995 和 2001 年 178 个职业的数据。这一数据来自于每年 3 月的"当期人口调查:年度人口档案"(Current population survey: Annual demographic file[CPS])。在 CPS 原始数据中,观察对象是个人,但我使用的只是 178 个职业的汇总数据。对于每个年度的每个职业,我都计算了该职业中女性的比例及女性中位工资收入。这里只考虑了每个年度至少有 50 个样本个体的 178 个职业。更详细的信息可以参看英格兰、埃里森及吴的作品(England, Allison & Wu, 2007)。工资变量被标定为 MDWGF1-MDWGF4,而女性比例为 PF1-PF4。

对于等式 6.4 中的模型,令 y 为中位工资收入,x 为女性比例。在 1983 年,这两个变量之间的相关系数为-0.33,并且高度显著。关于这两个变量之间的因果方向存在相当多的争论(England et al., 2007)。一种观点认为,雇主会贬低女性比例较高的职业的价值,从而支付较低的工资。与此针锋相对的假设是,逐渐下降的工资使得该职业对男性不再那么具有吸引力,当他们离开并涌向报酬更好的工作时,妇女将填补他们空缺下来的职位。这里,我假定两个变量中任何一个的变化都会在 6 年后另一个变量的变化中显露出来。

通过估计等式 6.4 中的两个方程，我们可以对这两种可能的因果作用进行评估。尽管这两个方程可以同时进行估计，但分开估计可使模型的设置具有更大的弹性。[31]除固定效应外，容许相互作用的关键设置在于：各个时点的误差项被允许与时变协变量的未来值(future value of time-dependent covariate)相关(Woodridge, 2002)。在我们的例子里，Time2 时中位工资方程中的误差项必须被允许与 Time3 时的女性比例之间存在非零的相关。与此类似，Time2 时女性比例方程中的误差项必须与 Time3 时的中位工资相关。注意：没有方程可以用来预测 Time1 时的中位工资或者女性比例，因为我们没有观察它们 6 年前(1977 年)的滞后值。

另外要注意，对于滞后因变量，只允许潜变量与 Time1 时的变量值相关。这是因为只有 Time1 变量是外生的，而相关只被允许存在于外生变量之间。事实上没有必要设定潜变量与滞后因变量后来的各取值相关，因为潜变量本身就是求取这些变量的方程的自变量之一。

两个方程的结果呈现在表 6.3 中。一点都不奇怪的是，每个变量都对其自身 6 年后(的取值)有着正向的、统计显著的作用。但对于“交叉—滞后”系数，没有证据支持两个作用方向中的任何一个。

表 6.3　交互作用模型的估计结果

自变量	反应变量			
	工资中位数		女性比例	
	系　数	标准误	系　数	标准误
工资中位数	0.344**	0.064	−0.001	0.002
女性比例	−0.159	2.447	0.299**	0.079

在其他地方，我曾经质疑过，当模型已经包含固定效应时，纳入因变量滞后值是否必要(desirability)(Allison, 1990)。因此，我还估计了一个不含因变量滞后值的模型，得到的交叉—滞后系数完全一样。类似的，含有因变量滞后值但不包括固定效应的模型(经典的两期—双变量面板模型)同样未能提供支持某一方向的交叉—滞后效应的证据。

总结

定量反应变量的线性固定效应或随机效应回归模型可以用 SEM 软件进行估计，所得结果与第 2 章中讨论的更常规的方法得到的一样。不过，这种方法要求不同的数据结构，这种数据的一条记录包括每个个体或群组的所有测量，多次测量被编码为不同的变量。在 SEM 软件中，每个反应变量在每个时点都设定了一个单独的方程，不同方程的系数通常被限定彼此相等。而随机效应或固定效应被设定为潜变量，并为各个方程中所共有。在固定效应形式下，这一潜变量被允许与所有在不同方程中有所变化的自变量之间存在相关。

这一方法通常比第 2 章中描述的方法设置起来更加麻烦。但是，它允许进行一些有趣的拓展，包括对固定效应和随机效应进行比较的似然比检验，对固定效应和随机效应模型的折衷，以及建立潜变量有多个指标的模型。最重要的是，在 SEM 框架下有可能对这样一种跟踪调查数据模型进行估计，在这种模型中，两个或多个反应变量彼此之间被认为存在滞后的相互作用。这种模型使我们有可能根据非实验数据做出比平常更有力的因果推论。

附录 1 | 第 2 章到第 5 章例题的 Stata 程序

```
use "C:\\data\\nlsy", clear

/*表 2.1*/
reg anti90 self90 pov90
reg anti94 self94 pov94
gen antidiff = anti94 - anti90
gen povdiff = pov94 - pov90
gen selfdiff = self94 - self90
reg antidiff povdiff selfdiff

/*表 2.2*/
reg antidiff povdiff selfdiff pov90 self90 black ///
  hispanic childage married gender momage momwork

/*表 2.3*/
gen antidif1 = anti92 - anti90
gen antidif2 = anti94 - anti92
gen selfdif1 = self92 - self90
gen selfdif2 = self94 - self92
gen povdif1 = pov92 - pov90
gen povdif2 = pov94 - pov92
```

```
reg antidif2 selfdif2 povdif2
reg antidif1 selfdif1 povdif1
gen id = _n
reshape long antidif povdif selfdif i(id)
gen eqdum = _j - 1
reg antidif povdif selfdif
xtset id _j
xtreg antidif povdif selfdif eqdum, pa

/* creat data set with 3 records per person */
use "C:\\data\\nlsy", clear
gen id = _n
reshape long anti self pov, i(id)
gen time = 1 + (_j - 90)/2

/* 表 2.5 */
xi: reg anti self pov i.time i.id
xi: reg anti self pov i.time
xtset id time
xi: xtreg anti self pov i.time, fe

/* 表 2.6 */
xi: xtreg anti i.time*self i.time*pov ///
  i.time*gender i.time*childage ///
  i.time*hispanic i.time*back i.time*momwork ///
  i.time*married ///
  i.time*momage, fe i(id)
testparm _ItimXself* _ItimXpov* _ItimXgend* ///
```

```
  _ItimXchill * _ItimXhisp * ///
  _ItimXblac * _ItimXmomw * _ItimXmarr * _ItimXmoma *

/ * 表 2.7 * /
xi: xtreg anti self pov i.time black hispanic ///
  childage married gender momage momwork
xi:xtreg anti self pov i.time

/ * Hausman test * /
xi: xtreg anti self pov gender childage hispanic ///
  black momwork married momage i.time
estimates store random_effects

xi: xtreg anti self pov i.time, fe
estimates store fixed_effects
hausman fixed_effects random_effects

/ * 表 2.8 * /
egen mself = mean(self), by(id)
egen mpov = mean(pov), by(id)
gen dself = self - mself
gen dpov = pov - mpov
xi: xtreg anti dself dpov mself mpov black ///
  hispanic childage married ///
  gender momage momwork i.time
test (dself = mself) (dpov = mpov)
xi: xtmixed anti dself dpov mself mpov black ///
  hispanic childage married ///
```

```
  gender momage momwork i.time || id: dself

/*表 3.1*/
use "C:\\data\\teenpov.dta", clear
tab pov1 pov5

/*表 3.2*/
drop if pov1 = = pov5
gen dmother = mother5 - mother1
gen dspouse = spouse5 - spouse1
gen dschool = inschool5 - inschool1
gen dhours = hours5 - hours1
logit pov5 dmother dspouse dschool dhours
logit pov5 dmother dspouse dschool dhours black age
logit pov5 dmother dspouse dschool dhours black ///
  age mother1 spouse1 inschool1 hours1

/*表 3.4*/
use "c:\\data\\teenpov.dta", clear
reshape long pov mother spouse inschool hours, i(id)
rename inschool school
rename _j year
xtset id year
xi: xtlogit pov mother spouse school hours ///
  i.year, fe
xi: xtlogit pov mother spouse school hours ///
  i.year, pa corr(uns)
xi: xtlogit pov mother spouse school hours i.year
```

```
/*表 3.5*/
gen mothblack = mother * black
xi: xtlogit pov mother spouse school hours ///
  mothblack i.year, fe
gen yearschool = (year - 1) * school
gen yearhours = (year - 1) * (hours - 8.67)
gen yearblack = (year - 1) * black
gen yearage = (year - 1) * (age - 15.65)
xi:xtlogit pov mother spouse school hours year ///
  yearschool yearhours ///
  yearblack yearage, fe

/*表 3.6, 3.7*/
egen mmother = mean(mother), by(id)
egen mspouse = mean(spouse), by(id)
egen mschool = mean(school), by(id)
egen mhours = mean(hours), by(id)
gen dmother = mother - mmother
gen dspouse = spouse - mspouse
gen dschool = school - mschool
gen dhours = hours - mhours
xi: xtlogit pov dmother dspouse dschool dhours ///
  mmother mspouse mschool mhours black age i.year
test dmother = mmother
test dspouse = mspouse
test dschool = mschool
test dhours = mhours
test (dmother = mmother)(dspouse = mspouse)///
```

```
  (dschool = mschool)(dhours = mhours)
xi: xtmelogit pov dmother dspouse dschool dhours ///
  mmother mspouse ///
  mschool mhours black age i.year ||id: dmother

/*表 3.8*/
use "C:\\data\\nlsy", clear
gen id = _n
reshape long anti self pov, i(id)
gen time = 1 + (_j - 90)/2
egen mself = mean(self), by(id)
egen mpov = mean(pov), by(id)
gen dself = self - mself
gen dpov = pov - mpov
xi: ologit anti dself dpov mself mpov black ///
  hispanic childage married ///
  gender momage momwork i.time, cluster(id)
test (dself = mself) (dpov = mpov)

/*表 3.9*/
use "C:\\data\\teenpov2.dta", clear
reshape long mother spouse empstat, i(id)
drop if empstat = .
gen currage = age + _j - 1
egen mmother = mean(mother), by(id)
egen mspouse = mean(spouse), by(id)
egen mage = mean(currage), by(id)
gen dmother = mother - mmother
```

```
gen dspouse = spouse - mspouse
gen dage = currage - mage
mlogit empstat dmother mmother dspouse mspouse ///
  dage mage black, ///
  vce(cluster id) base(1)
test([#1]dmother = [#1]mmother) ///
  ([#1]dspouse = [#1]mspouse)([#1]dage = [#1]mage)
test([#2]dmother = [#2]mmother) ///
  ([#2]dspouse = [#2]mspouse)([#2]dage = [#2]mage)
preserve
drop if empstat = 3
gen empstat2 = empstat - 1
xtset id _j
xtlogit empstat2 dmother ///
  mmother dspouse mspouse dage mage black, re
drop if empstat = 2
gen empstat3 = empstat - 1
xtset id _j
xtlogit empstat3 dmother mmother dspouse mspouse ///
dage mage black, re

/*表 4.1*/
use patents, clear
gen total = pat75 + pat79
gen rd0 = logr79 - logr75
gen rd1 = logr78 - logr74
gen rd2 = logr77 - logr73
gen rd3 = logr76 - logr72
```

```
gen rd4 = logr75 - logr71
gen rd5 = logr74 - logr70

blogit pat79 total
blogit pat79 total, vce(jack)
blogit pat79 total, vce(boot)
blogit pat79 total rd0 - rd5
blogit pat79 total rd0 - rd5, vce(jack)
blogit pat79 total rd0 - rd5, vce(boot)
blogit pat79 total rd0 - rd5 science logsize
blogit pat79 total rd0 - rd5 science logsize, vce(boot)

/*表 4.2*/
use patents, clear
rename pat75 patent1
rename pat76 patent2
rename pat77 patent3
rename pat78 patent4
rename pat79 patent5
gen sumpat = patent1 + patent2 + patent3 + patent4 + patent5
gen rda1 = logr75
gen rda2 = logr76
gen rda3 = logr77
gen rda4 = logr78
gen rda5 = logr79
gen rdb1 = logr74
gen rdb2 = logr75
gen rdb3 = logr76
```

```
gen rdb4 = logr77
gen rdb5 = logr78
gen rdc1 = logr73
gen rdc2 = logr74
gen rdc3 = logr75
gen rdc4 = logr76
gen rdc5 = logr77
gen rdd1 = logr72
gen rdd2 = logr73
gen rdd3 = logr74
gen rdd4 = logr75
gen rdd5 = logr76
gen rde1 = logr71
gen rde2 = logr72
gen rde3 = logr73
gen rde4 = logr74
gen rde5 = logr75
gen rdf1 = logr70
gen rdf2 = logr71
gen rdf3 = logr72
gen rdf4 = logr73
gen rdf5 = logr74
gen id = _n
reshape long patent rda rdb rdc rdd rde ref, i(id)
rename _j time
rename rda rd0
rename rdb rd1
rename rdc rd2
```

```
rename rdd rd3
rename rde rd4
rename rdf rd5
list id time patent rd0 - rd5 in 1/20
xtset id time

/* 表 4.3 */
xi: xtpoisson patent rd0 rd1 rd2 rd3 rd4 rd5 ///
  i.time, fe
xi: xtpoisson patent rd0 rd1 rd2 rd3 rd4 rd5 ///
  i.time, fe vce(boot)
xi: xtpoisson patent rd0 rd1 rd2 rd3 rd4 rd5 ///
  i.time, re
xi: xtpoisson patent rd0 rd1 rd2 rd3 rd4 rd5 ///
  i.time, pa corr(uns) vce(robust)
xi: xtpoisson patent rd0 rd1 rd2 rd3 rd4 rd5 ///
  i.time

/* 表 4.4 */
gen xdsci = rd0 * science
xi: xtpoisson patent rd0 rdsci i.time, fe
xit: xtpoisson patent rd0 rdsci i.time, fe vce(boot)

/* 表 4.5 */
gen scitime = time * science
xtpoisson patent rd0 time scitime, fe i(id)
xtpoisson patent rd0 time scitime, fe i(id) ///
  vce(boot)
```

```
/*表4.6*/
drop if sumpat = 0
xi: nbreg patent i.id rd0 rd1 rd2 rd3 rd4 rd5 ///
  i.time
xi:nbreg patent i.id rd0 - rd5 i.time,vce(opg)

/*表4.7*/
xi:xtnbreg patent rd0 - rd5 i.time, fe i(id)
xi:xtnbreg patent rd0 - rd5 science logsize ///
  i.time, fe i(id)

/*表4.8*/
egen mrd0 = mean(rd0), by(id)
egen mrd1 = mean(rd1), by(id)
egen mrd2 = mean(rd2), by(id)
egen mrd3 = mean(rd3), by(id)
egen mrd4 = mean(rd4), by(id)
egen mrd5 = mean(rd5), by(id)
gen drd0 = rd0 - mrd0
gen drd1 = rd1 - mrd1
gen drd2 = rd2 - mrd2
gen drd3 = rd3 - mrd3
gen drd4 = rd4 - mrd4
gen drd5 = rd5 - mrd5
xi: xtnbreg patent drd0 drd1 drd2 drd3 drd4 drd5 ///
  mrd0 mrd1 mrd2 mrd3 ///
  mrd4 mrd5 science logsize i.time, re
test (drd0 = mrd0) (drd1 = mrd1) (drd2 = mrd2) ///
```

```
  (drd3 = mrd3) (drd4 = mrd4) (drd5 = mrd5)
xi: xtnbreg patent drd0 drd1 drd2 drd3 drd4 drd5 ///
  mrd0 mrd1 mrd2 mrd3 ///
  mrd4 mrd5 science logsize i.time, pa robust
test (drd0 = mrd0) (drd1 = mrd1) (drd2 = mrd2) ///
  (drd3 = mrd3) (drd4 = mrd4) (drd5 = mrd5)

/*表5.1*/
use "C:\\data\\nsfg.dta", clear
stset dur, failure(birth = 1)
stcox pregordr age married passt nobreast lbw ///
  caesar multiple college, nohr
stcox pregordr age married passt nobreast lbw ///
  caesar multiple college, nohr cluster(caseid)

/*表5.2*/
stcox pregorder age married passt nobreast lbw ///
  caesar multiple college, ///
  strata(caseid) nohr
gen collbreast = college * nobreast
stcox pregorder age married passt nobreast lbw ///
  caesar multiple college ///
  collbreast, nohr strata(caseid)

/*表5.3*/
use "C:\\data\\coupleday.dta", clear
xtset coupleid day
xtlogit husdead wifed15, fe or
```

```
xtlogit husdead wifed30, fe or
xtlogit husdead wifed60, fe or
xtlogit husdead wifed90, fe or
xtlogit husdead wifed120, or
xtlogit husdead wifed15, or
xtlogit husdead wifed30, or
xtlogit husdead wifed60, or
xtlogit husdead wifed90, or
xtlogit husdead wifed120, or

/*表5.5*/
drop if wifefirst = 0
gen day2 = day * day
xtlogit wifed15 husdead day day2, fe or
xtlogit wifed30 husdead day day2, fe or
xtlogit wifed60 husdead day day2, fe or
xtlogit wifed90 husdead day day2, fe or
xtlogit wifed120 husdead day day2, fe or
```

附录 2 | 第 6 章例题的 Mplus 程序

```
! 表 6.1
! Random Effects
Data: file is "c:\\data\\nlsy.dat";
variable: names are anti90 anti92 anti94 black childage gender hispanic
married momage momwork pov90 pov92 pov94 self90 self92 self94; usevari-
ables = anti90 anti92 anti94 black childage gender hispanic married mo-
mage momwork pov90 pov92 pov94 self90 self92 self94;
Model:
  falpha by anti90 - anti94@1;
  anti90 on
    pov90 (1)
    self90 (2)
    black (3)
    hispanic (4)
    childage (5)
    married (6)
    gender (7)
    momage (8)
    momwork (9);
  anti92 on
    pov92 (1)
```

```
    self92 (2)
    black (3)
    hispanic (4)
    childage (5)
    married (6)
    gender (7)
    momage (8)
    momwork (9);
  anti94 on
    pov94 (1)
    self94 (2)
    black (3)
    hispanic (4)
    childage (5)
    married (6)
    gender (7)
    momage (8)
    momwork (9);
falpha with pov90 - pov94@0 self90 - self94@0 black@0 hispanic@0
childage@0 married@0 gender@0 momage@0 momwork@0;
anti90 anti92 anti94 (10);

! Fixed Effects
Data: file is "C:\\data\\nsly.dat";
Variable: names are anti90 anti92 anti94 black childage gender hispanic
married momage momwork pov90 pov92 pov94 self90 self92 self94; usevari-
ables = anti90 anti92 anti94 black childage gender hispanic married mo-
mage momwork pov90 pov92 pov94 self90 self92 self94;
```

```
Model:
  falpha by anti90 - anti94@1;
  anti90 on
    pov90 (1)
    self90 (2)
    black (3)
    hispanic (4)
    childage (5)
    married (6)
    gender (7)
    momage (8)
    momwork (9);
  anti92 on
    pov92 (1)
    self92 (2)
    black (3)
    hispanic (4)
    childage (5)
    married (6)
    gender (7)
    momage (8)
    momwork (9);
  anti94 on
    pov94 (1)
    self94 (2)
    black (3)
    hispanic (4)
    childage (5)
```

```
    married (6)
    gender (7)
    momage (8)
    momwork (9);
falpha with black@0 hispanic@0 childage@0 married@0 gender@0 momage
@0 momwork@0;
anti90 anti92 anti94 (10);

! Compromise
Data: file is "c:\\data\\nlsy.dat";
variable: names are anti90 anti92 anti94 black childage gender hispanic
married momage momwork pov90 pov92 pov94 self90 self92 self94; usevari-
ables= anti90 anti92 anti94 black childage gender hispanic married mo-
mage momwork pov90 pov92 pov94 self90 self92 self94;
Model:
  falpha by anti90 - anti94@1;
  anti90 on
    pov90 (1)
    self90 (2)
    black (3)
    hispanic (4)
    childage (5)
    married (6)
    gender (7)
    momage (8)
    momwork (9);
  anti92 on
    pov92 (1)
```

```
    self92 (2)
    black (3)
    hispanic (4)
    childage (5)
    married (6)
    gender (7)
    momage (8)
    momwork (9);
  anti94 on
    pov94 (1)
    self94 (2)
    black (3)
    hispanic (4)
    childage (5)
    married (6)
    gender (7)
    momage (8)
    momwork (9);
falpha with self90 - self94@0 black@0 hispanic@0 childage@0
    married@0 gender@0 momage@0 momwork@0;
anti90 anti92 anti94 (10);

! 表 6.3
Data: file is "C:\\data\\occ.dat";
Variable: names are pf1 - pf4 mdwgf1 - mdwgf4; usevariables pf1 - pf4 md-
wgf1 - mdwgf3;
Model:
  alpha by pf2 - pf4@1;
```

```
  pf4 on
    pf3 (1)
    mdwgf3 (2);
  pf3 on
    pf2 (1)
    mdwgf2 (2);
  pf2 on
    pf1 (1)
    mdwgf1 (2);
mdwgf3 with pf2;

Data: file is "C:\\data\\occ.dat";
Variable: names are pf1 - pf4 mdwgf1 - mdwgf4; usevariables pf1 - pf3 md-
wgf1 - mdwgf4;
Model:
  alpha by mdwgf2 - mdwgf4@1;
  mdwgf4 on
    pf3 (1)
    mdwgf3 (2);
  mdwgf3 on
    pf2 (1)
    mdwgf2 (2);
  mdwgf2 on
    pf1 (1)
    mdwgf1 (2);
mdwgf2 with pf3;
```

注释

[1] 感谢皮特·苔丝(Peter Tice)为我准备并提供了这一数据。

[2] 因为 ANTI 只能取整数值并且是正偏态分布,因此使用序次 logit 回归可能比使用线性回归更合适。事实上,在第 3 章我们就会用这种模型来分析该数据。不过,logit 模型得到的结论与本章利用线性模型得到的结论在性质上几乎是一样的。

[3] 在这里,随机效应模型(xtreg 命令的默认设置)是不合适的,因为随机效应模型只允许误差项之间存在正向相关。在存在多个一阶差分方程的情况下,误差相关通常是负的。

[4] 这里 reg 命令(本章讨论的所有其他 Stata 命令)是和 xi 前缀一起使用的,这样可以将 TIME 及 ID 变量作为分类变量处理。

[5] 这些数字可以通过将各变量分别作为因变量,ID 变量作为分类自变量进行方差分析得到。

[6] Stata 另一个实现离均差法的命令是用 areg 配合 absorb(id)选项来实现的。

[7] Hausman 检验的计算如下:令 b 表示固定效应系数向量(不包括截距项),β 表示对应随机系数向量。令 $\sum = \text{var}(b) - \text{var}(\beta)$,其中 var($b$)是 b 的估计协方差矩阵,β 的类似。Hausman 检验的统计量为 $m = (b-\beta)'\sum^{-1}(b-\beta)$,它在虚无假设下服从卡方分布。

[8] 这里描述的混合法与 Mundlak(1978)和 Hausman(1978)提出的方法相似,但并不相同。

[9] 这些估计只有在数据集平衡——也就是说,每一个个体被观察的期数相等——的情况下,才完全相等。否则,混合法所得估计将与常规固定效应估计略有差异。

[10] 关于这一结果的来历,请看埃里森的另一本著作(Allison, 2005)。

[11] 在两期观察情况下,条件似然法也可以采用下一节将讨论的针对三期及以上数据的方法进行,使用的是 Stata 中的 xtlogit 命令。所得结果将与刚刚讨论的"差分"法一样。

[12] 我同样拟合了另一个模型——YEAR 在全部 4 个交互项中都被当做定类变量对待,但通过似然比检验发现,该模型与模型 2 并不存在显著差异。

[13] 在表 3.5 中我仅仅考察了这些被选中的交互项。在很多实际应用中,可能需要同时检验所有变量与时间的交互项,以对模型在各时期的稳定性进行全面检验。这可以通过比较两个模型的方式完成:一个模型含有所有的交互项,而另一模型不含任何此类交互项。两个模型的似然比卡方统计量的差值本身就是对所有这些交互项系数都等于 0(相应的自由度等于两个模型的自由度之差)的假设的似然比卡方检验。

[14] 这些检验通过 Stata 中的 test 命令能够轻松完成。具体细节请看附录 1。

[15] 本章不考虑零膨胀泊松及负二项模型,有三个原因:它们要复杂得多,几乎没有软件可用来对跟踪调查数据进行此种分析,而且负二项模型本身就能为含有大量零计数值的数据提供满意的拟合。

[16] Stata 中的 reshape 命令使得数据结构的此种变换变得非常简单。

[17] 由于自助法带有随机抽样环节,因此每次运行所得自助标准误会略有差异。通过增加自助样本的数量可以使变异的程度变小。

[18] 表 4.3 中的 GEE 估计是使用 5 个年份专利数的"非结构化(unstructured)"相关矩阵所得结果。

[19] nbreg 命令能够拟合两种不同形式的负二项模型。在默认形式(这里比较适合采用的形式)下,方差是均值的函数。而在另外一种形式下,使用 dispersion(constant)选项可以将方差设置成为一个常数。这尽管看起来很具吸引力,但并不适合用在这里。

[20] 和第 3 章的 logistic 模型不同,负二项回归的总体均值模型与具体单位模型之间似乎不存在任何

差异。这意味着，随机效应估计结果从大小上看不应该比 GEE 估计更大。

[21] 实际上，我确实尝试这么做了，但我的电脑在运算了 10 天后仍在运算，我只好终止了这一尝试。从原则上讲，此种计算困难可以使用格林(Greene)的算法(2001)解决，但目前的商业软件中没有现成程序可用。

[22] 非常感谢尼古拉斯・克里斯塔吉斯(Nicholas Christakis)允许我在这里使用这些数据，对这一数据更详细的描述请看埃里森和克里斯塔吉斯的著作(Allison & Christakis, 2006)。

[23] 可以纳入时间的非单调函数，如 $\sin(2\pi t/365)$，它能够在一年当中周期性地变动。

[24] 当模型在控制变量上"饱和"时，这种对称是完全的，而对于非饱和模型，这种对称只是近似的。所谓饱和模型，是指含有分类自变量及其所有可能交互项的模型。

[25] 在 Stata 中使用用户提交的 gllamm 命令可以估计一些结构方程模型。但即使是这一命令，它的设定也相当笨拙、复杂。

[26] 我没有在表 6.1 中报告 TIME 的系数，因为用 Mplus 估计的这三个截距与表 2.5 中的并不完全对应。在表 2.5 中 TIME2 的系数等于时间 2 时的截距与时间 1 时的截距之差。类似的，表 2.5 中 TIME3 的系数等于时间 3 时的截距与时间 1 时的截距之差。

[27] 阿尔纳斯和霍尔姆(Ejrnaes & Holm, 2006)错误地宣称传统的固定效应估计结果与 SEM 估计结果不同。事实上，这两种方法总是给出相同的结果。

[28] SEM 检验有 6 个自由度，Hausman 检验有 4 个，而混合法检验只有 2 个自由度。这是因为 SEM 检验允许 α 和 x 之间的协方差在三个时期各不相同，而另外两种方法内在地限定它们相同。Hausman 检验比混合法检验多两个自由度，是因为它同时检验了两个时间系数在随机效应模型与固定效应模型中是否相等。拟合固定效应模型时，通过限定 α 和 x 之间的协方差在各个时期相等，我们可以在 SEM 框架下得到一个自由度为 2 的检验。对于 NLSY 例子来说，这将得到一个自由度为 2 的卡方值，对应 p 值为 0.003，略微小于自由度为 6 的检验的 p 值。这相对于混合法检验来说要小得多，后者得到的卡方值为 9.86，对应自由度为 2，p 值为 0.007。

[29] 关于另外一种使用 IV 的方法，可以参见豪斯曼和泰勒的作品(Hausman & Taylor, 1981)。

[30] 这些方程之所以能够用 OLS 进行估计，是因为第一个方程中的两个 x 与两个 ε 都保持独立，而这是因为 x 只受早先时点的 ε 的(间接)影响。同样的原理适用于第二个方程。

[31] 如果两个方程同时进行估计，那么不管是作为因变量还是作为自变量，每个变量都必须用同样的方式进行表达。但是，如果要将它们分开进行估计，那我们可以(采用不同形式的表达)，例如将一个变量的对数形式作为因变量，而在其作为自变量时使用非对数形式。

参考文献

Abrevaya, J. (1997). The equivalence of two estimators of the fixed-effects logit model. *Economics Letters*, 55, 41—44.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1—10.

Allison, P. D. (1984). *Event history analysis: Regression for longitudinal event data*. Beverly Hills, CA: Sage.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. In C. Clogg (Ed.), *Sociological methodology 1990* (pp. 93—114). Oxford, UK: Basil Blackwell.

Allison, P. D. (1995). *Survival analysis using SAS*. Cary, NC: SAS Institule.

Allison, P. D. (1996). Fixed effects partial likelihood for repeated events. *Sociological Methods & Research*, 25, 207—222.

Allison, P. D. (1999a). *Logistic regression using SAS: Theory and application*. Cary, NC: SAS Instute.

Allison, P. D. (1999b). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge.

Allison, P. D. (2000, June). *Inferring causal order from panel data*. Paper prepared for presentation at the Ninth International Conference on Panel Data, Geneva, Switzerland.

Allison, P. D. (2002). *Bias in fixed-effects Cox regression with dummy variables*. Unpublished paper, Department of Sociology, University of Pennsylvania.

Allison, P. D. (2004). Convergence problems in logistic regression. In M. Altman, J. Gill, & M. McDonald(Eds.), *Numerical issues in statistical computing for the social scientist* (pp. 247—262). New York: Wiley InterScience.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Cary, NC: SAS Institute.

Allison, P. D., & Bollen, K. A. (1997, August). *Change score, fixed effects, and random component models: A structural equation approach*. Paper presented at the annual meeting of the American Sociological Association.

Allison, P. D., & Christakis, N. (2006). Fixed effects methods for the analysis of non-repeated events. In R. Stolzenberg(Ed.), *Sociological methodology 2006* (pp. 155—172). Oxford, UK: Basil Blackwell.

Allison, P. D., & Waterman, R. (2002). Fixed effects negative binomial regression models. In R. M. Stolzenberg(Ed.), *Sociological methodology 2002* (pp. 247—265). Oxford, UK: Basil Blackwell.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Ecornmic studies*, 58, 277—297.

Baltagi, B. H. (1995), *Econometric analysis of panel data*. New York: Wiley.

Begg, C. B., & Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, *71*, 11—18.

Bryk, A. S. , & Raudenbusch, S. W. (1992). *Hierarchical linear models: Application and data analysis methods*. Newbury Park, CA: Sage.

Cameron, A. C. , & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.

Center for Human Resource Research. (2002). *NLSY 97 user's guide*. Washington, DC: U. S. Department of Labor.

Chamberlain, G. A. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies*, 47, 225—238.

Chamberlain, G. A. (1985). Heterogeneity, omitted variable bias, and duration dependence. In J. J. Heckman & B. Singer(Eds.), *Longitudinal analysis of labor market data* (pp. 3—38). Cambridge, UK: Cambridge University Press.

Conaway, M. R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association*, *84*, 53—62.

Cox, D. R. (1972). Regression models and life tables(with discussion). *Journal of the Royal Statistical Society*, *Series B*, *34*, 187—220.

Darroch, J. N. , & McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics*, *28*, 371—388.

Dunteman, G. H. , & Ho, M. R. (2005). *An introduction to generalized linear models*. Thousand Oaks, CA: Sage.

Ejrnaes, M. , & Holm, A. (2006). Comparing fixed effects and covariance structure estimators for panel data. *Sociological Methods & Research*, 35, 61—83.

England, P. , Allison, P. D. , & Wu, Y. (2007). Does bad pay cause occupations to feminize, does feminization reduce pay, and how can we tell with longitudinal data? *Social Science Research*, *36*(3), 1237—1256.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Griffin.

Greene, W. H. (2000). *Econometric analysis*(4th ed.). Upper Saddle River, NJ: Prentice Hall.

Greene, W. H. (2001). *Estimating econometric models with fixed effects*. New York University, Leonard N. Stern School, Finance Department Working Paper Series.

Greenland, S. (1996). Confounding and exposure trends in case-crossover and case-time control designs. *Epidemiology*, *7*, 231—239.

Hall, B. H. , Griliches, Z. , & Hausman, J. A. (1986). Patents and R and D: Is there a lag? *International Economic Review*, *27*(2), 265—283.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. Cary, NC: SAS Institute.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, *46*(6), 1251—1271.

Hausman, J. A. , Hall, B. H. , & Griliches, Z. (1984). Econometric models for count data with an application to the patents-R & D relationship. *Econometrica*, *52*, 909—938.

Hausman, J. A. , & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, *49*, 1377—1398.

Honoré, B. E. (1993). Orthogonality conditions for tobit models with fixed effects and lagged

dependent variables. *Journal of Econometrics*, *59*, 35—61.

Honoré, B. E., & Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometricsa*, 68, 839—874.

Hsiao, C. (1986). *Analysis of panel data*. Cambridge, UK: Cambridge University Press.

Judge, G. Hill, C., Griffiths, W., & Lee, T. (1985). *The theory and practice of econometrics*. New York: Wiley.

Kalbfleisch, J. D., & Sprott, D. A. (1970). Applications of likelihood methods to models involving large numbers of parameters(with discussion). *Journal of the Royal Statistical Society*, *Series B*, *32*, 175—208.

Kenward, M. G., & Jones, B. (1991). The analysis of categorical data from cross-over trials using a latent variable model. *Statistics in Medicine*, *10*, 1607—1619.

Kline, R. B. (2004). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kreft, I. G. G., & De Leeuw, J. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1—21.

LaMotte, L. R. (1983). Fixed-, random-, and mixed-effects models, In S. Kotz, N. L. Johnson, & C. B. Read(Eds.), *Encyclopedia of statistical sciences*(pp. 137—141). New York: Wiley.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, *95*, 391—413.

Lewis-Beck, M. S. (1995). *Data analysis: An introduction*. Thousand Oaks, CA: Sage.

Long, J. S. (1983). *Covariance structure models: An introduction to LISREL*. Beverly Hills, CA: Sage.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Maclure, M. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, *133*, 144—153.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A non-parametric approach to statistical inference*. Newbury Park, CA: Sage.

Mundlak, Y. (1978). On the pooling of time series and cross sectional data. *Econometrica*, *56*, 69—86.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376—398.

Muthén, B., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *2*, 371—402.

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, *54*(2), 638—645.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.

Senn, S. (1993). *Cross-over trials in clinical research*. New York: Wiley.

Singer, J. B., & Willett, J. D. (2003). *Applied longitudinal data analysis: Modeling change and*

event occurrence. New York: Oxford University Press.

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel(Eds.), *Handbook of statistical modeling for the social and behavioral sciences*(pp. 1—38). New York: Plenum Press.

Suissa, S. (1995). The case-time-control design. *Epidemiology*, *6*, 248—253.

Teachman, J., Duncan, G., Yeung, J., & Levy, D. (2001). Covariance structure models for fixed and random effects. *Sociological Methods and Research*, *30*, 271—288.

Therneau, T. M., & Grambsch, P. (2000). *Modeling survival data: Extending the Cox model*. New York: Springer-Verlag.

Tjur, T. A. (1982). Connection between Rasch item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, *9*, 23—30.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.

译名对照表

average treatment effects	平均处理效应
between R^2	组间确定系数
bootstrap standard errors	自举标准误
case-crossover method	案例-交叉法
case-time-control method	案例-时间-控制法
censored cases	删失案例
censored intervals	删失区间
conditional maximum likelihood	条件最大似然法
constant variance assumption	恒定方差假定
convergence failure	收敛失败
cross-lagged coefficients	交叉滞后系数
deviance statistic	离差统计量
deviation coefficients	离差变量系数
difference scores	差分值
duration analysis	存活期分析
dynamic models	动态模型
endogenous variables	内生变量
event history analysis	事件史分析
exogenous variables	外生变量
failure time analysis	失效时间分析
first difference equation	一阶差分方程
first difference method	一阶差分法
frailty term	脆弱成分
generalized estimating equations	广义估计方程(GEE)法
generalized least squares(GLS) regression	广义最小二乘回归
gompertz model	Gompertz 模型
group mean centering	组均值对中
hausman test	Hausman 检验
hazard analysis	风险分析
incidental parameters problem	伴随性参数问题
instrumental variables(IV)	工具变量(IV)框架
jackknife standard errors	刀切法标准误
lagged dependent variable	滞后因变量
latent variables	潜变量模型
likelihood ratio test	似然比检验

linear structural equation model	线性结构方程模型
log-linear model	对数线性模型
monte carlo simulation	蒙特卡罗模拟
NB2 model	负二项模型
odds ratios	发生比率
ordered logit model	次序 logit 模型
overdispersion	过离散
panel models	面板模型
panel survey	固定样本跟踪调查
partial likelihood method	偏似然法
path diagrams	通径图
poisson models for count data	计数数据泊松模型
proportional hazards model	比例风险模型
random effects models	随机效应模型
random intercept models	随机截距模型
random slope models	随机斜率模型
robust standar errors	稳健标准误
saturated model	饱和模型
shared frailty models	共享脆弱性模型
stratification	分层
strictly exogenous variable	严格外生变量
structural equation model(SEM)	结构方程模型(SEM)
subject specific coefficient	具体单位系数
survival analysis	生存分析
time-invariant variables	非时变变量
unconditional maximum likelihood	无条件最大似然法
vector of coefficients	系数向量
within R^2	(组、个体)内确定系数
zero-inflated Poisson models	零膨胀泊松模型
zero mean	零均值

现代稳健回归方法

作者简介

罗伯特·安德森

罗伯特·安德森(Robert Anderson)是加拿大多伦多大学社会学和政治科学教授。他的研究兴趣是应用统计学、政治社会学(尤其是态度及政治行为的社会基础)、社会分层和工作社会学(the sociology of work)。曾在 *American Sociology Review*, *The Journal of politics* 和 *Sociological Methodology* 等期刊发表论文。

译者简介

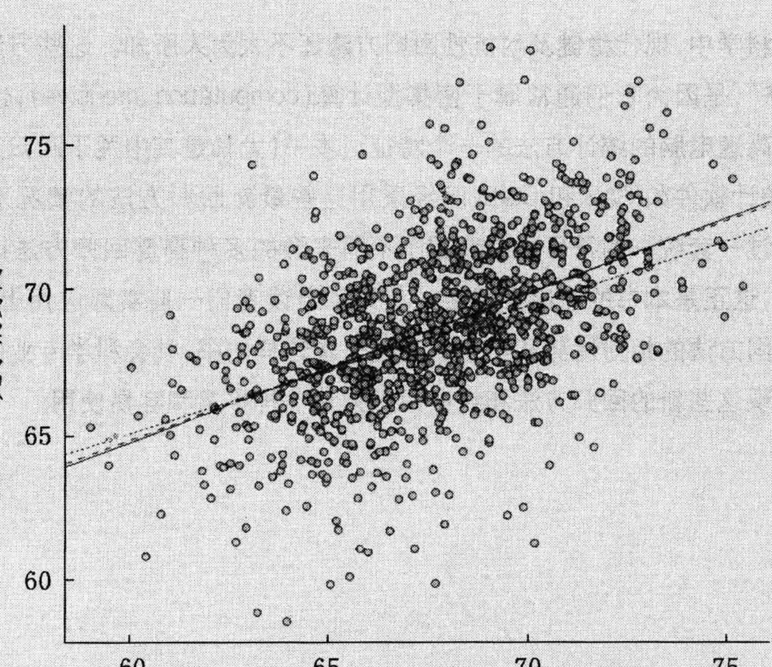
李丁

李丁,北京大学社会学系本科和硕士毕业,现为北京大学社会学系博士研究生。

序

1886 年，弗朗西斯·高尔顿（Francis Galton）发表了题为《遗传身高向普通回归》（Regression Towads Mediocrity in Hereditary Stature）的开创性文章，从而开启了今天我们所知的线性回归统计方法的发展历程。通过分析 205 对父母及 928 个小孩的数据，高尔顿发现相对较高或较矮的父母生养的小孩倾向于不是那么高或矮，这一特征被统计术语概括为"向均值回归"。

为了演示回归是如何处理此类身高数据的，我使用了一套相似但只有一个性别的数据，这应归功于高尔顿的徒弟卡尔·皮尔森（Karl Pearson）。下图标绘出了 1078 对父子的身高状况（单位是英寸），数据用小圈点表示，它们明显地遵循一种线性趋势，刻画出向均值（等于 45 英寸）回归的现象。



在本图中，我拟合了一条回归直线，用实线表示，斜率估计值为 0.514，由一般最小

二乘估计得到(这一估计及以后其他估计的双尾检验都比常规的0.001水平显著得多,因此这里就不报告了)。不管以谁的标准来看,这一数据的表现都很不错。不过,即使是在这一表现良好的数据里面,有些案例也比其他的更异常:我们很快就可看到图中右上角及左下区的某些案例离其他围绕在直线周边的大多数案例更远。如果这些案例太过极端,我们就可以从下列标准的快速“处理办法”中选择一个:从分析中剔除这些案例、重新编码(如果存在编码错误的话),以及在分析中纳入更多新变量。但如果没有处理这些异常(或不那么异常)案例的合理可用的解决办法,数据分析者该怎么办呢?这正是稳健及耐抗性回归方法(robust and resistant regression method)派得上用场的地方。

为了展示一下稳健回归,我对上述数据拟合了另外两条直线(使用的是R软件里的MASS数据包),虚线表示的是用MM—估计量(MM-estimator)估计得到的稳健回归线(斜率估计值 = 0.502),点线表示的是通过将分位残差平方最小化(minimization of quantile squared residuals)的耐抗性回归估计(估计过程中分位残差最大的案例被忽略)得到的直线(斜率 = 0.442)。可以看到,使用MM—估计得到的稳健回归结果,其斜率只比OLS回归的稍小。不过,耐抗性回归得到的估计结果差别更大,所得出的结论表现出更严重的向均值的回归。由安德森撰写的这本著作的焦点在于有效性(validity)的(而非效率[efficiency]的)稳健,它将帮助社会科学研究者理解这些方法,并学到稳健回归的原理及应用方法。

在社会科学中,现代稳健及耐抗性回归方法还不太为人所知。这些方法之所以被称为“现代方法”,是因为它们通常属于密集型计算(computation intensive),这是当前很多依赖今天的高速电脑的统计方法的一个特征。本书(尤其是其中关于回归方法的那些章节)在主要统计软件如SAS和Stata已经采用这些最新回归方法的情况下是非常及时的。本书通过一套统一的符号系统介绍了不同来源的多种稳健回归方法以及它们彼此之间的联系,这正是本书的杰出贡献之一。为了给读者们一些实际应用上的帮助,本书也讨论了不同方法的相对优势和不足。通过一本这样的书,社会科学专业的学生及研究者最终会发现这些新的回归方法和经典回归方法一样平常和容易使用。

廖福挺

第1章 | 绪言

在定量社会科学中，回归分析是统计方法的主要干将。大量的问题都是由线性模型或者广义线性模型（generalized linear model）解决的。只要被恰当使用，回归估计能够为数据里的关系提供有效而简洁的概括。但如果盲目而机械地使用，回归分析也会导致错误的结论。异常观察案例的存在就是引起担心的原因之一，它们有时足以严重扭曲由一般最小二乘（OLS）回归所估计的结果，哪怕数据集很大。异常观察值也能对广义线性模型造成破坏性的损害，虽然这不是很常见。这进一步强化了发现并恰当地处理回归分析中的特异值/异常值（outlier）的重要意义。

将“现代”回归方法，如非参数回归，作为诊断工具整合进一般线性模型及广义线性模型的框架，有很多好处（参见，例如 Cook and Weisberg，1999；Fox，1997；Hastie，Tibshirani and Friedman 2001）。这些方法之所以被称为“现代方法”，是因为它们依赖于密集计算，即在拟合大量回归的基础上计算出最终估计结果，它们能够揭示出只使用 OLS 估计时常常难以发现的大量问题——尤其是非线性问题，当然也包括其他残差方面的问题。只是在个人电脑运算速度已经极大提高的最近这段时间，社会统计学家才意识到这些方法的好处。

稳健回归是另一套接近密集计算型的现代技术。当我们面对特异值束手无策时——也就是说，它们不能被重新编码，也无法通过进行形式变换或在模型中纳入新因素的方式得到处理时——稳健回归就是 OLS 回归之外的合适选择。很多类型的稳健回归有着共同的目标，即提供不受特异值或偏态残差分布影响的无偏估计。这些方法最令人满意的地方在于：当误差项服从正态分布时，它们也会努力提供具有相对效率的估计。稳健回归技术至少还能作为探测

潜在问题案例的有用的诊断工具(diagnostic tool)。

虽然稳健回归在社会科学中应用得不广，但统计学家已经知道它的好处几十年了，并且一直在发展新的方法。最近，有几本出色的专著面向统计学家们讨论了这些方法(参见，如 Atkinson and Riani, 2000; Lawrence and Arthur, 1990; Myers, 1990; Maronna, Martin, and Yohai, 2006; Wilcox, 2005)。不过，除了一些关于稳健回归的一般性讨论文章外(如 Western, 1995; Dietz, Frey and Kaloff, 1987; Wu, 1985)，社会科学实践者们还很难找到一本对于众多不同稳健回归的差异进行探讨的专著。而您面前的这本书正试图改变这一状况。

本书试图讨论各种用来探测及恰当处理回归分析中的“权势案例”(influential cases)①的方法。第 2 章将界定一些对于理解估计的稳健性至关重要的术语。由于位置(location)与尺度(scale)的各种测度，构成了稳健回归技术的基础，因此这一章也对它们进行了讨论。第 3 章列出了异常观察案例及偏态分布影响 OLS 估计的不同方式。同时也简要介绍了一般线性回归中探测权势案例的一些传统技术——包括正式的统计检验和绘图法(graphical methods)。第 4 章讨论了各种线性模型稳健回归方法及其限制。第 5 章则讨论了稳健回归估计的标准误，主要集中在自助法(bootstrapping)上。第 6 章简要描述了广义线性模型和在这种模型中探测异常观察案例的几种诊断法。更重要的是，这一章将稳健回归方法扩展到了广义线性模型。第 7 章将对全书进行总结，并给出如何处理异常观察值的一些一般建议。最后，附录部分对常用电脑统计软件中的稳健回归工具进行了总结。

全书的编排时刻考虑到实际的研究过程：一旦我们建立起某个模型以检验自己的假设，那么怎样做才能确保不会出现问题案例而使得我们的推论和检验变得不准确呢？再有，如果此类问题案例真的存在，我们该如何处理呢？在解答这些大问题之前，对稳健性——更确切地讲，强健(robust)且耐抗(resistant)的回归——的一般概念进行定义，并为本书提供一个具有启发性的真实案例非常重要。

---

① 由于“影响”的汉语词义限制，在本书中表示观察案例对回归估计影响的 influence 或 influential 被互译为权势、影响或影响力等，以使不同表达更为顺畅。——译者注

何为“稳健”(Robustness)?

如休伯(Huber, 2004:1)所言,“‘稳健’一词负载了很多有时并不完全一致的含义”。尽管如此,一般都认为对一个估计量(estimator)进行评价时,需要考虑两种类型的稳健性。这些是由莫斯特勒和图基(Mosteller & Tukey, 1977:203—209)总结的,他们认为一个稳健的估计量必须满足两个条件:(1)数据的微小改动,将不会造成估计的剧烈变化;(2)在各种情况下该估计都具有高度的效率。第一个条件,反映的是估计量抵抗特异值干扰的抗扰性/耐抗性(resistance),可以被看作效度的稳健性(robustness of validity)。换句话说,该估计量为数据的主体部分提供了有效的估计。第二个条件,与潜在的分布假定(underlying distributional assumption)有关,可以被当作是效率的稳健性(robustness of efficiency)。这一条件意味着,估计量的分布假定未被满足对其精度影响很小(换言之,对其标准误的影响很小)。

本书主要关注效度的稳健性(robustness of validity)——也就是说,一个估计量在遇到异常观察值时保持不变的水平。当然,也会讨论效率的稳健性(robustness of efficiency),尽管对于一个稳健估计量来讲,它只被看作是一个第二位的标准。因此,在描述了估计量效度上的稳健性后,我们将会对其效率进行一定的讨论。必须澄清的是,尽管有偏分布(skewed distribution)和特异值/离群点(outlier)在概念上存在差异,但二者对于一个估计量可以造成类似的后果,不管是效度上的稳健性还是效率上的稳健性。不过,本书主要讨论的还是特异值的影响。简言之,本书对试图抑制异常观察案例影响的多种估计量的稳健性进行了比较。从这个意义上讲,抗扰性和稳健性在整本书中被当作同义词使用——它们都被用来指示单个观察案例对估计量的影响程度。

稳健回归(Robust Regression)的定义

并不奇怪的是,对于回归分析而言,“稳健”也负载着很多含义。定义之一

与所谓的“稳健标准误”(robust standard errors)相关,后者常被用来处理某些模式的异方差问题(heteroscedasticity)或误差相关(error dependency)问题。这一含义虽然很有用,但并不适合本书关于稳健回归的界定。前面已经说过,本书讨论的是各种专门试图适应——说得更好点,减轻——异常观察案例(影响)的方法。就这一点而言,稳健回归有两种定义与此直接相关。

第一种将所有明确适应重尾型误差分布(heavy-tailed error distributions)及特异值的回归模型都叫做稳健回归模型。第二种区分了稳健回归(robust regression)和耐抗回归(resistant regression)。根据这种定义,稳健回归技术同时与效度的稳健性及效率的稳健性相关。这些技术用的是来自所有观察案例的信息,但给那些高度异常案例的权重较轻。很多稳健回归都考虑了残差异常值,自变量上的特异值(杠杆效应,leverage),或者两者的组合。这些方法中的大多数都能给出有效率的估计,不管误差分布是重尾分布还是正态分布。

相反,通常被称为耐抗回归的方法一般很少考虑效率问题。这些方法的主要目标,在于防止异常观察值剧烈影响回归斜率的取值。它们不仅降低了异常观察案例的权重,并且经常设立一定的标准,以将特异案例完全剔除出分析。

人们通常还会将稳健回归区分为崩溃点/失效点高的(with a high breakdown point)与崩溃点低的(with a low breakdown point)。另外一种分类方式是根据模型是否存在有界限的影响(bounded influence)进行的。这些概念将在第2章中详细定义;就目前而言,只要知道一个非常稳健的估计量应该具备受限制的影响和高崩溃点就够了。这些区别曾经非常重要,不过最新发展的稳健回归已经将高崩溃点和有限影响结合起来了。

因此,就本书的目的而言,稳健回归被宽泛地定义为任何限制特异值对回归估计造成过分影响的回归。因此,这些不同的回归技术不会被分为“稳健”类或“耐抗”类。相反,本书正是从各种方法对已有方法做了怎样的发展这一角度来对它们进行探索的,当然,上述所有标准在各种方法的讨论中都会涉及。

一个真实的例子：20世纪70年代已婚夫妇的性生活频率

也许大家都知道，异常观察案例在小样本情况下会给回归估计带来问题，而且这一点从直观上就能理解。如果 n 很小，用以平衡高度异常案例的个案就少。相反，数据很大时，异常案例要严重改变回归平面会难很多，因为有大量观察案例可以平衡它们。以简单回归为例，我们知道，回归直线是通过将残差平方和 $\sum e_i^2$ 最小化而得到的，其中 e_i 即 y_i 的观测值与回归直线预测值 $\hat{y}_i$ 之差 ($e_i = y_i - \hat{y}_i$)。如果在几千个观测值中只有一个极端值，那它要想将回归直线拉向自己非常困难，因为还有如此众多的与之对抗的残差需要维持在很小的规模。当然，这并不是说异常案例在大数据中不能对回归估计造成严重破坏。下面的例子就能很好地说明它们如何可能造成此种影响。

加索(Jasso, 1985)使用全国生育调查(National Fertility Studies)跟踪数据(panel data)研究了在控制队列效应的情况下，年龄与时期对1970—1975年期间已婚夫妇每月性生活频率的影响。她的主要发现为：(1)在控制了队列和年龄效应后，月经周期影响为负；(2)控制月经周期和队列效应后，妻子的年龄有着正向作用。这些发现与以往的研究结论大不相同，因此不奇怪的是，文章在《美国社会学评论》(*American Sociology Review*)发表了。

值得赞扬的是，加索对她所用的方法讨论得非常清楚，因此其他研究者可以重现她的结果。卡恩和尤德利(Kahn & Udry, 1986)正好做了这样的重现工作，并对加索的原初分析提出了几点疑虑。首先，他们认为有4个案例似乎被错误地编码成了88。他们认为这些取值实际上应该被编码为99，而这在该数据集中是缺失信息的赋值。他们得出这一结论是因为没有其他的受访者报告过超过63的取值，且99.5%的观测值小于40。其次，通过使用模型诊断技术，卡恩和尤德利还发现了另外4个特异值。这4名受访者在前一轮调查中报告的次数要少很多，这意味着这些特异案例并不典型，可以正当地从模型中移除。最后，他们认为加索没有加入婚龄长度与妻子年龄之间的交互项。卡恩和尤德利对

这一数据的再分析——剔除了这 8 个特异值(从超过 2000 个案例的样本中)并加上交互项——完全改变了原来的结论。如表 1.1 所示，在特异值被移除后，加索的新发现看起来就远远没有那么重要了。更具体地说，妻子的年龄(考虑到非线性模式的存在，已经取了对数)在标准水平上已经不再统计显著了。

表 1.1 性生活频率的决定因素

	模型 1	模型 2	模型 3	模型 4
月经周期	−0.72***	−0.67***	−3.06**	−0.08
妻子年龄的对数	27.61**	13.56	29.49	−1.62
丈夫年龄的对数	−6.43	7.87	57.89	−5.23
婚龄长度的对数	−1.50***	−1.56***	−1.51*	1.29
妻子怀孕状况	−3.71	−3.74***	−2.88***	−3.95*
有 6 岁以下小孩	−0.56**	−0.68***	−2.91***	−0.55**
妻子处于在职状态	0.37	0.23	0.86	0.02
丈夫处于在职状态	−1.28**	−1.10**	−4.11***	−0.38
R^2	0.0475	0.0612	0.2172	0.0411
n	2062	2055	243	1812

资料来源：引自 Kahn and Udry(1986，表 1)。

*$p<0.05$；**$p<0.01$；***$p<0.001$。

注：模型 1：Jasso 的原始分析。模型 2：4 个“错编”案例和另外 4 个特异值删除后；模型 3：婚龄小于或等于 2 年(已排除错编及极端案例)；模型 4：婚龄大于等于 2 年(已排除错编及极端案例)。

加索在她的回应(Jasso，1986)中认为：卡恩和尤德利的分析将产生“样本删截偏差(sample truncation bias)”的新问题。她表示，通过移除极端值，卡恩和尤德利人为地将因变量局限在了其值域的一个小片段内。她反对这一做法，并主张研究者不应该因为异常观察案例未能服从一已信念就将它们从样本中删除。她还表示，对于一个大规模的全国性随机样本来说，选出几个差异较大的观察案例是完全可能的，因为根据文化和地区的不同，夫妻性生活频率的差异很大。

这个例子向我们说明了三个要点。第一，它展示了通过使用诊断工具发现潜在问题案例所具有的价值。第二，它展示了在大样本中极端值是如何影响回归估计的。第三，其中的争论，说明并不存在被普遍接受的异常案例处理办法。

发现存在权势案例后，决定采取何种动作，必须有丰富的知识作为基础。换言之，研究者必须凭借自己的判断力做出决定。至于这个具体的例子，我把决定哪一种特异值处理方式最好的机会留给那些在这个话题上实际知识更加丰富的人。就本书的目的而言，上面的讨论已经足够了，它充分地说明：即使是在一个超过 2000 个案例的样本中，仅仅只有 8 个特异值就能如此剧烈地改变研究的结果。如果有人认定这些案例不宜忽视，那他也可以像卡恩和尤德利一样直接删除它们，或者采用稳健回归。

第 2 章 | 重要背景

接下来我们讨论对于评估一个估计量的稳健性非常重要的各种概念。在此，偏差(bias)、一致性(consistency)、崩溃点/失效点(breakdown point)以及影响函数(influence function)等概念将会得到定义。而且这些概念都将自始至终贯穿全书。

偏差与一致性(Bias and Consistency)

假定样本 Z 有 n 个案例。令 $T_n(Z_1, \cdots, Z_n)$为参数 θ 的一个估计量，且其概率分布为 P。换句话说，将 T 运用于 Z 即得到总体参数的估计值：

$$T(Z) = \hat{\theta} \qquad [2.1]$$

一个无偏估计量需满足：

$$E[T(Z)] = E(\hat{\theta}) = \theta \qquad [2.2]$$

也就是说，无偏统计量的平均值等于总体的参数。由此可以推出，一个估计量 $T(Z) = \hat{\theta}$ 的偏差通过下式可以得到：

$$\text{bias } E[T(Z) - \theta] \qquad [2.3]$$

选用一个“最好”的统计量，无偏性无疑非常重要，但一致性也需要考虑。如果随着样本规模增大估计量 $\hat{\theta}$ 向 θ 收敛，那它就具备一致性。我们也可从均方误(mean squared error, MSE)的角度来理解估计量的一致性。就此而言，满足如下条件的 $\hat{\theta}$ 就具备一致性：

$$\lim_{n \to \infty} \text{MSE}(\hat{\theta}) = 0 \qquad [2.4]$$

崩溃点/失效点(Breakdown Point)

崩溃点(BDP)[1]是一种估计量的抗异常值干扰能力的全局性测度(global measure of the resistance)。准确地讲,它是一个估计量在不产生任意结果(arbitrary result)的前提下能够容忍的离群案例(如特异值或聚集在分布尾端尽头的案例)的最小分量或百分比(Hampel, 1974; Huber, 2004)。假定所有的可能"破败(corrupted)"样本——其中有 m 个观察案例被替换为任意值(即不符合数据一般趋势的观察案例)——为 Z'。那么,由这种替换所可能造成的最大影响[2]是:

$$\text{effect}(m;\ T,\ Z) = \sup_{Z'} \| T(Z') - T(Z) \| \qquad [2.5]$$

其中上确界(supremum)是所有可能的 Z' 样本数。如果 effect(m; T, Z)无限(infinite),那么这 m 个特异值对 T 有任意大的影响。换句话说,该估计量"崩溃"了,不能充分代表数据主体部分的模式。更一般地讲,一个有限样本 Z 的估计量 T 的崩溃点是这样定义的:

$$BDP(T,\ Z) = \min\left\{\frac{m}{n}:\ \text{effect}(m;\ T,\ Z)\ \text{无限}\right\} \qquad [2.6]$$

一个估计量最高的可能崩溃点是 50%,也就是说多达一半的观察案例可以被忽视(discounted)。高于 0.5 的崩溃点是不可取的,因为它意味着该估计量仅仅与一小部分数据相关。

稳健估计量的目标在于充分地把握数据的主体模式。换句话说,崩溃点大于零就算一种好的属性。实际上,汉普尔等人(Hampel et al., 1986)认为:一个数据集通常会有多达 10%的案例偏离数据主体部分所具有的一般模式,这意味着稳健估计量的崩溃点至少要达到 10%。不过,后面我们即将看到,有些最早提出的稳健回归估计量的崩溃点为 0 或者非常接近于 0。

影响函数(Influence Function)

估计量的影响函数最先由汉普尔(Hampel，1974;又见 Hoaglin，Mosteller and Tukey，1983:350—358; Jurečková and Picek，2006:27—32)提出，它衡量的是对估计量 T 的理论假定分布 F 造成污染的单个观察案例 y_i 的影响(impact)。换句话说，如果崩溃点衡量的是全局稳健性(global robustness)，那么影响函数(IF)衡量的就是局部抗扰性(local resistance)，更确切地讲，对估计量的极微小扰动(infinitesimal perturbations)。影响函数通常也被称为影响曲线(infuence curve,从单个样本来看，又称为灵敏度曲线，即 sensitivity curve)，估计量 T 的影响函数是这样定义的：

$$IF(Y,\ F,\ T) = \lim_{\lambda \to 0} \frac{T[(1-\lambda)F + \lambda\delta_y] - T(F)}{\lambda} \qquad [2.7]$$

其中 δ_y 是在 y 点(即在 y,或者 0 点)的概率质量为 λ 的污染度(the point of contamination at y)。换言之,λ 给出了在 y 点的污染比例。简单地讲,IF 表示在 y 点增加任意特异值所引起的估计量的变化,这一变化经过了损害比例标准化。

有界的影响函数(bounded influence function)是稳健估计量应该具备的属性之一,因为它意味着一个案例的影响最多也就达到该高度。一个没有限制/边界的影响函数允许“被污染”观察值的影响持续增长,不管它们有多么异常。也就是说,偏差值的影响根本没有界限。后面我们将会看到,OLS 回归的影响函数是没有边界的,且与残差的大小成比例,这意味着偏差极大的残差能够完全毁坏 OLS 估计量。很多早期稳健回归方法的影响函数也是没有边界的,结果其抗扰性有时不比 OLS 好。不过,今天所用的大多数稳健回归都具有高崩溃点和有界影响函数。

相对效率(Relative Efficiency)

另一个对于理解稳健估计非常重要的概念是效率。如果目标是要从样本

数据来对一个更大的总体进行推论，那么我们就需要有一个效率尽可能高的无偏统计量。从严格意义上讲，估计量的效率取决于它可能的最小方差与实际方差的比率。只有当这一比率等于1时——也就是说，当它的方差等于可能的最小方差时——一个估计才被认为是有效率的。[3]如果一个估计量能够在大样本时达到有效率，就被认为是渐近有效率的(asymptotically efficient)。更一般地讲，抽样方差相对较小，从而标准误也较小的估计量被认为是高效率的。由此可以推知，有些估计量会比另外一些效率更高，因此，相对效率(relative efficiency)的概念在对各种估计量进行评估比较时非常有用。

对于大多数种类的估计，总会有一个估计量在一定的假定条件下效率最高。我们可以用这个估计量作为比较其他估计量的效率的标准。假定对于总体参数 θ 我们有两个估计量 T_1 和 T_2。如果 T_1 的效率最高，T_2 较差，那么 T_1 的均方误将相对较小。T_2 的相对效率由它的均方误与 T_1 的均方误的比率决定：

$$\text{Effiviency}(T_1,\ T_2) = \frac{E[(T_2-\theta)^2]}{E[(T_1-\theta)^2]} \qquad [2.8]$$

如果线性假定(linearity)、误差方差恒定(constant error variance)及误差项独立(uncorrelated errors)这几个假定都满足的话，那么OLS估计就是最有效率的无偏线性估计。因此，稳健估计量的相对效率就是通过在这些条件下与OLS估计量进行比较得到的。尽管在这些条件下没有哪个稳健回归估计的效率比OLS回归更高，但是一些估计量不仅具有极其相近的效率，而且同时具备抵抗特异值干扰的良好属性。不过，稳健回归估计的相对效率必须被小心地对待，因为通常所评定的效率仅仅是渐近效率而已(Ryan, 1997:354)。也就是说，相对效率只有在样本规模足够大时才有意义。多数稳健估计量在样本规模很小时的属性尚不太清楚，因此，在这种情况下，人们通常会使用自助法来获得标准误。

位置测度/位置量数(Measures of Location)

尽管存在各种各样的回归，但它们通过自变量来预测因变量的条件取值时，都考虑了反应变量本身位置及尺度(scale)的测度。例如，OLS就是从一个

或多个自变量 x 来估计因变量 y 的条件均值的。由于 OLS 建立在不具有抗特异值干扰能力的平均值基础上，因此它的估计同样会被特异值影响。类似的，来自广义线性模型(generalized linear models)的估计也不能完全抵抗特异值干扰，因为它们估计的是某种线性预测值(linear predictor)的条件均值。稳健回归方法依赖于更稳健的位置和(或)尺度测度。因此，在探讨运用了这些位置和尺度测度的回归技术之前，对这些测度本身进行讨论会有好处。

位置的测度即对分布中的某个位置进行刻画的量(quantity)。最典型的，对分布中心的测量就非常重要，当然其他位置测度(如，分位数)也可以被考虑。假定随机变量 Y 的分布为 F。如果对于任意常量 a 和 b，估计量 $\theta(Y)$ 满足下面四个条件[4]，那它就是 F 的位置测度之一(Wilcox, 2005:20—21)：

a. $\theta(Y+a)=\theta(Y)+a$

b. $\theta(-Y)=-\theta(Y)$

c. $Y\geqslant\theta$ 意味着 $\theta(Y)\geqslant 0$

d. $\theta(bY)=b\theta(Y)$

条件(a)要求所有的 Y 值都加上一个常量后，位置测度将增加同样的量，这被称为位置同变性(location equivariance)。条件(a)、(b)、(c)合在一起，要求该测度的取值在 Y 的值域范围内。而条件(d)意味着该测量需具有尺度同变性(scale equivariance)。换句话说，如果所有 Y 值都乘以一个特定的值(换言之，尺度发生变化)，那么位置的测度也会发生同等比例的变化。

均值(the Mean)

最常见的位置测度就是均值。假设有独立观察值 y_i 和一个估计总体分布中心 μ 的简单模型：

$$y_i=\mu+e_i \qquad [2.9]$$

其中 e_i 代表残差。如果潜在分布为正态分布，那么样本均值是效率最高的 μ 的估计量，得到的拟合模型为：

$$y_i = \bar{y} + e_i \qquad [2.10]$$

尽管均值使用非常广泛(包括在 OLS 回归中),但它并不是一个稳健的位置测度。如果该分布为重尾分布或者存在特异值,那均值比其他很多中心测度(mearsures of center)的效率都低,更重要的是,(这种情况下)它经常具有误导性。哪怕只增加一个被严重错误编码的案例,都能改变它的估计。

试看下面 5 个 y 变量的观察值:

$$y_1 = 3 \quad y_2 = 3 \quad y_3 = 4 \quad y_4 = 5 \quad y_5 = 5$$

通过使用大家都知道的样本均值计算公式 $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,可得到 $\bar{y} = 4$。现在将其中的一个观察值,比如 y_3,替换为一个"坏"值(假定它是错误编码值),得到各 y 值如下:

$$y_1 = 3 \quad y_2 = 3 \quad y_3 = 44 \quad y_4 = 5 \quad y_5 = 5$$

对于这一新数据而言,$\bar{y} = (3+3+44+5+5)/5 = 12$。这一均值被严重地拉向特异值,它是特异值被排除时的 3 倍。实际上,这个"被污染"均值比"坏"观测值之外的任何观察值都要大很多。

因为仅仅一个观察案例就可以使得均值崩溃失效,因此它的崩溃点是 $BDP = \frac{1}{n}$,当 n 很大时,这个取值在效果上相当于是 0。同样成问题的是,每个案例对均值的影响与 y 的大小成比例。而均值是通过将最小二乘目标方程(least squares objective function)最小化得到的:

$$\sum_{i=1}^{n} (y_i - \hat{\mu})^2 = 0 \qquad [2.11]$$

对 y 求导得到影响函数如下:

$$IF_{\bar{y}}(y) = 2y \qquad [2.12]$$

显然,这对一个并非"表现很好(well behaved)"(即带特异值或长尾)的数据来说,不是什么好的属性。

消除特异值对于均值造成影响的策略之一是使用一种两步程序(two-step procedure):先将特异值甄别出来,然后在计算均值前将它们删掉。汉普尔(Hampel, 1974)认为:使用一种稳健的位置测度往往比两步法更好。换言之,很多估计量比均值更加稳健。

α-截尾均值(α-Trimmed Mean)

一个相对稳健的中心测量(measure of center)是截尾均值(trimmed mean),这种均值通过删除分布尾端的观察值来降低特异值或重尾的影响。令 y_1, …, y_n 表示来自一个随机样本的某变量的观察值。我们先将 y 值从小到大顺序排列,$y(1) \leqslant y(2) \leqslant \cdots \leqslant y(n)$,并确定需要截除的量 $0 < \alpha < 0.5$,然后用排除了 $g(g = [\alpha n])$ 个最大值和最小值后的观察值计算出平均值,其中$[\alpha n]$取整为最近的整数值。截尾均值的计算公式可写成[5]:

$$y_t = \frac{y_{(g+1)} + \cdots + y_{(n-g)}}{n-2g} \qquad [2.13]$$

截尾均值的崩溃点取决于剪除量,因此 $BDP = \alpha$。一种简单的拇指法则是从分布的每个尾端各移除 10% 的观察案例(即设 $\alpha = 0.2$)。莱杰和罗马诺(Leger and Romano, 1990)进一步建议分别计算 $\alpha = 0$, 0.1 及 0.2 时的均值,并从中选出标准误最小的那个值作为最终的计算值。剪除量同样决定着影响函数。与均值不同,截尾均值的影响(函数)是有界的,尽管在 y_α 和 $y_{1-\alpha}$ 处有着明显的增长[6]。它的影响函数可以写成:

$$IF_{\bar{y}_t}(y) = \begin{cases} \dfrac{y_\alpha - \hat{\mu}_t}{1-2\alpha} & \text{当 } y < y_\alpha \text{ 时} \\ \dfrac{y - \hat{\mu}_t}{1-2\alpha} & \text{当 } y_\alpha \leqslant y \leqslant y_{1-\alpha} \text{ 时} \\ \dfrac{y_{1-\alpha} - \hat{\mu}_t}{1-2\alpha} & \text{当 } y > y_{1-\alpha} \text{ 时} \end{cases} \qquad [2.14]$$

其中 $\hat{\mu}_t$ 为截尾均值(见 Wilcox, 2005:29)。截尾均值的相对效率取决于分布情况。如果分布是正态的而裁剪过量,那么精度会降低,因为它会导致相对于较

小的 n 来说太大的宽度(spread),进而扩大了对抽样分布宽度(即标准误)的估计。相反,如果该分布带有重尾或特异值,截尾能够提高效率,因为 y 的方差——进而其均值的抽样分布的方差——减小了。究竟截尾占多大比例,只有在对分布做了仔细检查之后才能决定。

中位数(The Median)

中位数 M,即把数据从小到大排列时出现在中间位置的 y 的取值。要找到中位数,首先得将观察值按照从小到大的顺序排列 $y(1) \leqslant y(2) \leqslant \cdots \leqslant y(n)$。中位数即:

$M = y_{((n+1)/2)}$(当 n 是奇数时)

$M = 0.5y_{(n/2)} + 0.5y_{(n/2+1)}$(当 n 是偶数时)

同样的,中位数也得最小化目标函数的绝对值:

$$\sum_{i=1}^{n} | y_i - \hat{\mu} | = 0 \qquad [2.15]$$

取方程 2.15 的导数,即得到影响函数的形状

$$IF_M(y) = \begin{cases} 1 & 当\ y > 0\ 时 \\ 0 & 当\ y = 0\ 时 \\ -1 & 当\ y < 0\ 时 \end{cases} \qquad [2.16]$$

如这一受限的影响函数所示,中位数对特异值有着高度耐抗性。它的稳健性从它的 $BDP = 0.5$ 也能反映出来。中位数的不足之处在于,当分布为正态分布时它的效率要比均值低。在这种情况下,均值的抽样方差是 δ^2/n,而中位数的抽样方差是 $\pi\delta^2/2n$,是前者的 $\pi/2 = 1.57$ 倍。

尺度测度(Measures of Scale)

令 Y 代表一个随机变量。所谓尺度测度,是满足下列条件的任意非负函数 $\tau(Y)$(Wilcox, 2005:34)[7]:

a. 该测量是尺度同变(scale equivariant)的，意即 $\tau(aY) = a\tau(Y)$，其中 a 是大于 0 的任意常数。

b. 该测量是位置不变(*location invariant*)的，意即 $\tau(Y+b) = \tau(Y)$，其中 b 是常数。

c. 该测量是符号不变(sign invariant)的，$\tau(Y) = \tau(-Y)$

尺度测度实在太多，无法在这里全部囊括进来，因此，我们集中讨论那些与稳健回归最相关的测度。我们将主要探讨特异值如何影响尺度估计的大小(magnitude)，而很少关注效率问题。关于后者的更多讨论请看 Wilcox，2005。

标准差

使用最多的尺度测度是标准差 s，它是这样定义的：

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \qquad [2.17]$$

如果 y 的分布是正态的，这就是一个最合适的尺度测度，因为它具有极高的效率。但是，标准差在重尾分布或者带有特异值的分布面前并不稳健。因为它是基于均值的——后者的影响函数没有上限且崩溃点为 0——标准差继承了这些特性。因此，稳健回归技术通常使用其他尺度测度。

平均离差(Mean Deviation From the Mean)

相对于均值的平均离差(MD)有时被简称为平均离差(Mean Deviation)，它是这样定义的：

$$MD = \frac{\sum_{i=1}^{n}|y_i - \bar{y}|}{n} \qquad [2.18]$$

当 y 的分布带有长尾时，与标准差相比，MD 具有相对效率，但它仍然具有崩溃点为 0 以及影响函数没有界限的不良性质。尽管在一些早期稳健回归技术

中很重要，不过在目前存在众多更为稳健的尺度测量的情况下，MD 一般被认为已经过时了。

相对于中位数的平均离差(Mean Deviation From The Median)

相对于中位数的平均离差，即 MDM，在稳健性上要比 MD 有所改进。MDM 不是计算 y 与均值的绝对差异，而是计算它们与中位数 M 的绝对距离，结果为：

$$MDM = \frac{\sum_{i=1}^{n} |y_i - M|}{n} \qquad [2.19]$$

尽管 MDM 使用的是中位数，但它仍然依赖于平均离差，因此它的崩溃点 $BDP = 0$ 而且影响函数没有界限(具体请参见 Wilcox，2005:35)。也就是说，相对于中位数的平均离差对于极端特异值和长尾并不免疫，因此它并不是用于稳健回归的理想尺度测量。

四分位差(Interquartile Range)

q-分位差 QR_q 是一组影响有界限的尺度测量，它们的崩溃点非常高。任意特定的 q-分位差是这样给定的：

$$QR_q = y_{1-q} - y_q (\text{其中 } 0 < q < 0.5)$$

令 $q = 0.25$(也就是 0.25 分位与 0.75 分位之差)，即得到四分位差(IQR)，它的崩溃点 $BDP = 0.25$，它是最稳健因而也是最常用的分位差(Wilcox，2005:35—36)。四分位差的影响函数由 3/4 分位的影响函数减去 1/4 分位的影响函数(也就是 $IF_{0.75} - IF_{0.25}$)给定：

$$IF_{IQR}(y) = \begin{cases} \dfrac{1}{f(y_{0.25})} - C & \text{如果 } y < y_{0.25} \text{ 或 } y > y_{0.75} \\ -C & \text{如果 } y_{0.25} \leqslant y \leqslant y_{0.75} \end{cases} \qquad [2.20]$$

其中：

$$C = q\left\{\frac{1}{f(y_{0.25})} + \frac{1}{f(y_{0.75})}\right\} \qquad [2.21]$$

IQR 的崩溃点高并且影响函数有界，这些性质正是我们需要的，因此其在早期的稳健回归技术中有所使用。它在后面即将介绍的分位回归中扮演着一定角色。不过，尽管四分位差非常简单，但是因为还有更稳健的尺度测量，因此在新近发展的稳健回归中一般很少使用。

中位绝对离差(Median Absolute Deviation)

中位绝对离差(MAD)是这样定义的：

$$MAD = \text{median}\,|\,y_i - M\,|\ (\text{其中 } M \text{ 是中位数})$$

MAD 完全取决于围绕中位数的变异，因此比标准差及和均值相关的绝对离差抵抗特异值干扰的能力更强。[8] MAD 达到了最高的崩溃点 $BDP = 0.5$，它的有界影响是这样定义的：

$$IFMAD(y) = \frac{sign(|\,y - M\,| - MAD) - \dfrac{f(M+MAD) - f(M-MAD)}{f(M)}sign(y-M)}{2[f(M+MAD) + f(M-MAD)]} \qquad [2.22]$$

其中 $f(y)$ 是 y 的概率密度函数（更多内容见 Wilcox，2005：35）。MAD 一个诱人的特征在于它可以通过乘以 1.4826[近似于 $1/\Phi^{-1}(3/4)$，其中 Φ 是正态概率密度函数]而得到调整，从而保证在样本规模很大时的一致性。所有这些特征都使得 MAD 成为一个有吸引力的稳健回归的尺度测量，至少是作为一个初步估计。

M-估计(M-Estimation)

M-估计包括很多估计方法，它们将最大似然的思想推广用于尺度和位置的稳健测度（Huber，2004）。M 估计也是很多稳健回归估计的基础，包括那些被分类为 M 估计，GM 估计，S 估计以及 MM 估计的方法。这些都将在第 4 章得

到讨论。通过适当的公式表达，M 估计非常稳健，尤其是对于位置的估计。与其他用于大样本 ($n \geqslant 40$) 的稳健测量相比，它们也具有相对效率，并且随着 n 变大而更有效率(Hogg，1974;也见 Wu，1985)。

假定 y_1，…，y_n 相互独立并且都服从相同的分布 $F(y; \theta)$。令 $T_n(y_1, \cdots, y_n)$ 是刻画该分布 $F(y; \theta)$ 的未知参数 θ 的一个估计量。那么该估计量的似然值是这样定义的：

$$L(\theta; y_i, \cdots, y_n) = \prod_{i=1}^{n} f(y; \theta) \tag{2.23}$$

其中 $f(y; \theta)$ 是与 $F(y; \theta)$ 对应的概率密度函数。而最大似然估计结果是令似然函数最大化或(等价地)令目标函数 $\rho(y; \theta)$ 最小化的 θ 值：

$$-\log l = \sum_{i=1}^{n} \rho(y; \theta) \tag{2.24}$$

将目标函数限定为带有严格连续导数 $\Psi(\cdot)$ 的任何可微函数(function that is differentiable with an absolutely continuous derivative)，就得到最大似然估计量 T_n，

$$\sum_{i=1}^{n} \Psi(y; \theta) = 0 \tag{2.25}$$

其中：

$$\begin{aligned} \Psi(y; \theta) &= -(\partial/\partial\theta)\rho(y; \theta) \\ &= (\partial/\partial\theta)\log f(y; \theta) \end{aligned} \tag{2.26}$$

为了使最大似然估计——或者 M 估计——的解为唯一，$\rho(y; \theta)$ 必须为严格的凸函数，这样 $\Psi(y; \theta)$ 才会严格递增。使用 $\rho(y; \theta) = -\log f(y;\theta)$ 得到的是普通最大似然估计。(见 Huber，2004:第 3 章)

M 估计有很多不同的形式，其性质取决于对 $\rho(\cdot)$，或者与其等价的 $\Psi(\cdot)$ 的选择。如果 $\Psi(\cdot)$ 是没有边界的，那么估计量的崩溃点为 $BDP = \lim_{n\to\infty} BDP = 0$。相反，如果 $\Psi(\cdot)$ 是奇函数并且有界，那么 $\rho(\cdot)$ 以 0 为中心对称。得分函数(score function) $\Psi(\cdot)$ 的形状与汉普尔(Hampel，1974)提出的影响函数的形状

相同。更具体地说，$IF(y;\ F,\ T)=\Psi(y)/\gamma(F)$，其中 $\gamma(F)=\int f(y)d\Psi(y)$。比例常数 $[\gamma(F)]^{-1}$ 同时取决于 Ψ 和概率密度函数 $f(y)$。也就是说，IF 是负的得分函数(参见 Jureckova and Sen，1996；Hoaglin et al.，1983:356)。

位置的 M 估计(M-Estimation of Location)

如果总体的均值是随机变量 Y 的期望值。令 $\rho(y-\hat{\mu})$ 为测量相对于位置估计 $\hat{\mu}$ 的距离的目标函数。

$$\sum_{i=1}^{n}\rho(y;\ \theta)=\sum_{i=1}^{n}\rho\left(\frac{y_i-\hat{\mu}}{cS}\right) \qquad [2.27]$$

其中 S 是分布尺度的一个测量，c 为通过定义分布的中心和尾巴来对估计量的稳定程度进行调整的细调常数。尽管 M 估计具有位置同变性，但它们并不具备尺度同变性，因此细调常数是必需的。c 值越小，该估计抵抗特异值干扰的能力越大。

取等式 2.27 的导数，就得到影响函数的形状。M 估计值即能够解出下一等式的 $\hat{\mu}$ 的取值：

$$\sum_{i=1}^{n}\Psi\left(\frac{y-\hat{\mu}}{cS}\right)=0 \qquad [2.28]$$

尺度测量和位置测度是被同时估计的，因此需要一种迭代的估计程序(具体细节见 Huber，2004)。关于估计的更多细节将在第 5 章的回归的 M 估计中给出。在这里，我们接着讨论由均值延伸而来的一般解释(continue with a general explanation extending from mean)。

均值的 M 估计基于最小平方目标方程：

$$\rho(y;\ \theta)=\frac{1}{2}(y-\hat{\mu})^2 \qquad [2.29]$$

等式 2.28 的导数表明(案例的)影响与 y 值成比例：

$$\Psi(y;\ \theta)=(y-\hat{\mu}) \qquad [2.30]$$

为了计算一个比均值更稳健的 M 估计，我们只需要将最小平方目标方程替换为另外一个给极端值较小权重的方程即可。Huber 权重函数(Huber weight function)和双权数函数(biweight functions)是两种最常见的选择。

Huber 估计(Huber Estimates)

Huber 权重函数在分布的中心时表现得像均值和与之相连的最小平方目标函数(也就是说:观察案例被给予相同的权重)。但在分布的两端时，有点像中位数及与之相关的最小绝对值目标函数，它给分布尾部越靠外的观察值的权重越小：

$$\rho_{\mathrm{H}}(y;\ \theta)=\begin{cases}\frac{1}{2}y^2 & \text{如果 } y\leqslant c\\ c|y|-\frac{1}{2}c^2 & \text{如果 } y>c\end{cases}\qquad [2.31]$$

因为目标是产生一个能抵抗特异值干扰的估计，所以 MAD 通常被用来计算尺度测量 S。当总体服从正态分布时，定义 $S=MAD/0.6745$，得到 S 估计的 σ(S estimating σ)。遵照休伯(Huber, 1964)的建议，设定 $c=1.345$ 比较方便(统计软件中使用的也是这一标准)，它能够较好地抵抗特异值的干扰 ($1.345/0.6745 \cong 2$MADs) 并且相对效率接近 95%。

计算等式 2.31 的导数，得到影响函数的形状：

$$\Psi_{\mathrm{H}}(y;\ \theta)=\begin{cases}c & (\text{如果 } y>c)\\ y & (\text{如果 } y\leqslant |c|)\\ -c & (\text{如果 } y<-c)\end{cases}\qquad [2.32]$$

最后，$\Psi(\cdot)$的导数可以算出给予每个观察案例的权重：

$$w_{\mathrm{H}i}(y)=\begin{cases}1 & (\text{如果 } y\leqslant c)\\ c/|y| & (\text{如果 } y>c)\end{cases}\qquad [2.33]$$

双权数估计(Biweight Estimates)

双平方权重(Bisquare weight)，也被叫做 Tukey's 双平方，它与 Huber 权重

的主要区别出现在分布尾部的顶端，在这里双权数目标函数能更好地抵抗特异值的干扰。

$$\rho_{\mathrm{BW}}(y;\ \theta)\begin{cases}\dfrac{c^2}{6}\left\{1-\left[1-\left(\dfrac{y}{c}\right)^2\right]^3\right\} & (\text{如果}\ |y|\leqslant c)\\[2ex] \dfrac{c^2}{6} & (\text{如果}\ |y|>c)\end{cases} \qquad [2.34]$$

当细调常数 $c=4.685$ 时，$4.685\times S\simeq 7\mathrm{MAD}$，如果样本是从正态总体中抽取的，它能产生 95%的效率（Huber，1964）。计算等式 2.34 的导数，可以看到影响函数趋近于 0 的速度非常快。

$$\Psi_{\mathrm{BW}}(y;\ \theta)=\begin{cases}y\left[1-\left(\dfrac{y}{c}\right)^2\right]^2 & (\text{如果}\ |y|\leqslant c)\\[2ex] 0 & (\text{如果}\ |y|>c)\end{cases} \qquad [2.35]$$

取方程 2.35 的导数，得到权重函数：

$$w_{\mathrm{BW}\,i}(y)=\begin{cases}\left[1-\left(\dfrac{y}{c}\right)^2\right]^2 & (\text{如果}\ |y|\leqslant c)\\[2ex] 0 & (\text{如果}\ |y|>c)\end{cases} \qquad [2.36]$$

图 2.1 呈现了将带有默认细调常数的 Huber 函数和双权数函数应用于同一个值域范围为−10 到 10 的分布时的情形。可以看到，相对于与均值（它给所有的观察值同样的权重）的相似程度，两个 M 估计量的表现彼此更为相像。Huber

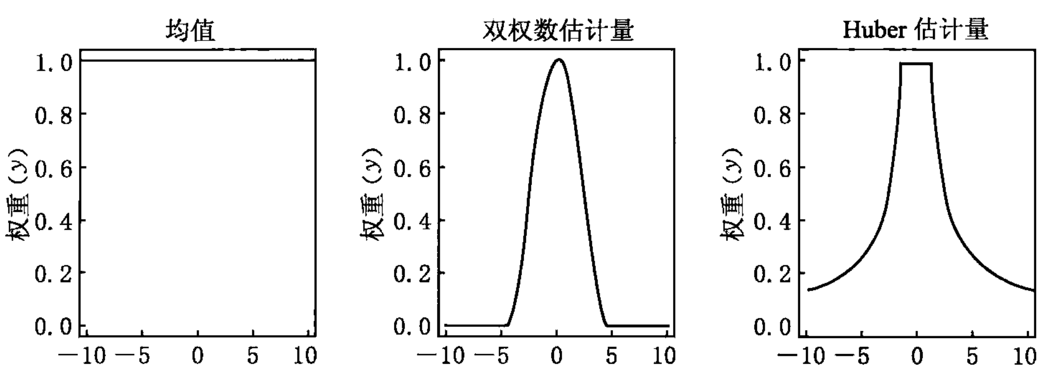


图 2.1　常用 M 估计量的权重函数与均值（的权重函数）的对比

表 2.1 几种可能的 M 估计函数

	目标方程,$\rho(\mu)$	影响函数,$\Psi(\mu)$	权重函数,$w(\mu)$
最小二乘	$\rho_{LS}(\mu)=\frac{1}{2}\mu^2$	$\Phi_{LS}(\mu)=\mu$	$w_{LS}(\mu)=1$
Huber	$\rho_H(\mu)=\begin{cases}\frac{1}{2}\mu^2 & 如果\mu\leqslant c\\ c\lvert\mu\rvert-\frac{1}{2}m^2 & 如果\mu>c\end{cases}$	$\Phi_H(\mu)=\begin{cases}c & 如果\mu>c\\ \mu & 如果\mu\leqslant c\\ -c & 如果\mu<-c\end{cases}$	$w_H(\mu)=\begin{cases}1 & 如果\mu\leqslant c\\ c/\lvert\mu\rvert & 如果\mu>c\end{cases}$
双权	$\rho_{BW}(\mu)=\begin{cases}\frac{c^2}{6}\left\{1-\left[1-\left(\frac{\mu}{c}\right)^2\right]^3\right\} & 如果\lvert\mu\rvert\leqslant c\\ \frac{c^2}{6} & 如果\mu>c\end{cases}$	$\Phi_{BW}(\mu)=\begin{cases}\mu\left[1-\left(\frac{\mu}{c}\right)^2\right]^2 & 如果\lvert\mu\rvert\leqslant c\\ 0 & 如果\lvert\mu\rvert>c\end{cases}$	$w_{BW}(\mu)=\begin{cases}\left[1-\left(\frac{\mu}{c}\right)^2\right]^2 & 如果\lvert\mu\rvert\leqslant c\\ 0 & 如果\lvert\mu\rvert>c\end{cases}$
Andrew	$\rho_A(\mu)=\begin{cases}c\{1-\cos(\mu/c)\} & 如果\lvert\mu\rvert\leqslant c\pi\\ 2c & 如果\lvert\mu\rvert>c\pi\end{cases}$	$\Phi_A(\mu)=\begin{cases}\sin(\mu/c) & 如果\lvert\mu\rvert\leqslant c\\ 0 & 如果\lvert\mu\rvert>c\end{cases}$	$w_A(\mu)=\begin{cases}\frac{\sin(\mu/c)}{\mu/c} & 如果\lvert\mu\rvert\leqslant c\\ 0 & 如果\lvert\mu\rvert>c\end{cases}$
Ramsay	$\rho_R(\mu)=\frac{1-e^{-c\lvert\mu\rvert(1+c\lvert\mu\rvert)}}{c^2}$	$\Phi_R(\mu)=\mu e^{-c\lvert\mu\rvert}$ (c^{-1} 处的最大值)	$w_R(\mu)=e^{-c\lvert\mu\rvert}$

资料来源:引自 Draper and Smith,1998:表 25.1。

和双权数函数在分布的绝大多数的部分作用近似，除了正中间和尾巴的末端(extreme tails)。双权数函数给所有绝对值大于 5($|y_i|>5$) 的观察值的权重为 0。相反，Huber 权重从不会让一个案例的权重为 0，而且会让更大比例的案例的权重等于 1。

尽管 Huber 权数函数及双权数函数是 M 估计中最常用的，不过还有很多其他选择，其中一些呈现在表 2.1 中。关于这些估计量的更多具体内容，尤其是相关细调常数的建议值，可以看安德鲁斯及其他人的作品(Andrews, et al., 1972)还有拉姆齐的文章(Ramsay, 1977)。

尺度的 M 估计量(M-estimator of scale)

将 M 估计扩展到对于尺度的估计相对直接(Wilcox, 2005:92—98)。同样的，主要的思想是找到一个给极端案例赋权较小的函数。尺度的 M 估计的一般形式是由位置的 M 估计的渐近方差定义的。

$$\zeta^2 = \frac{K_2\tau_2 E[\Psi^2(Z_i)]}{\{E[\Psi'(Z_i)]\}^2}$$
$$Z_i = \frac{y_i - \mu m}{cS} \qquad [2.37]$$

其中 μ_m 是位置的 M 估计，c 是一个正的细调常数，S 是通常被设为 MAD 的尺度的最初测量，Ψ 是得分函数。和位置的 M 估计一样，Huber 权数函数和双权数函数是典型的选择。因为它们使用得更为频繁而且已经表明效率更高，这里我们集中关注后者，它带来的是双权中位方差(biweight midvariance)(见 Lax, 1985)

双权中位方差抵抗特异值干扰的能力强，且很有效率，崩溃点接近 0.5(Hoaglin et al., 1983)。它是这样定义的：

$$\hat{\zeta}^2_{\text{bimid}} = \frac{\sum_{i:y_i^2\leqslant 1}(y_i - M_y)^2(1-Z_i^2)^4}{\sum_{i:y_i^2\leqslant 1}[(1-Z_i)^2(1-5Z_i^2)]^2} \qquad [2.38]$$

其中 M_y 是 y 的中位数，且：

$$Z_i = \frac{y_i - \mu_m}{cS} \qquad [2.39]$$

非常值得注意的是,等式中的求和被限定在 $y_i^2 \leqslant 1$ 的条件下。细调常数 c 通常设定为 9,尺度设定为 MAD,能够得到最大的效率。

各种估计的对比

例 2.1:仿真数据(Simulated Data)

表 2.2 对比了前面讨论过的一些估计量在用于仿真数据时的稳定性。第 1 列是将这些估计量用于来自标准正态分布 $y_i \sim N(0, 1)$ 的 20 个随机观测值(取值从−2.2 到 1.7)的情况。换句话说,这些数据都表现良好,没有特异值。第 2 列将这些估计量用于同样的数据,但增加了一个取值为 60 的极端特异值,假定它是被错误编码的案例。这些估计量的崩溃点放在了第 3 列。

表 2.2 含有极端特异值和不含极端特异值的仿真数据的位置及尺度测度

估计量	崩溃点	所有案例	特异值被移除
位置测度			
均值	0	0	2.85
α截尾均值	α(剪除比例)	−0.09	−0.04
中位数	0.5	−0.02	0.005
M估计	0.5	−0.12	−0.03
尺度测度			
标准差	0	1	13.13
平均离均差	0	0.71	5.44
平均离中差	0	0.61	2.89
四分位差	0.25	1.07	1.21
中位绝对离差	0.5	0.61	0.66
双权中位方差	0.5	0.89	1.06

表格的上半部分给出的是各种位置测度的结果。和 $BDP = 0$ 相一致,均值因为特异值的拉动而被严重扭曲(由 0 变为 2.85)。相反,截尾均值——按照常规,从尾部剪除了 20%的观察值,把特异值删除掉了——表现得非常好,它在好数据中和在受污染的数据中的值几乎完全一样(−0.09 比−0.04)。中位数及

M 估计(用的是双平方权重)的 *BDP* 都等于 0.5,二者看起来也没有受特异值影响。

再来看尺度测度,可以看到在运算与均值有关的——也即标准差、相对均值的平均离差、相对中位数的平均离差——都被特异值严重扭曲。当然,这毫不奇怪,因为它们的 *BDP* 都等于 0。标准差所受的影响最大,取值是不存在特异值时的 13 倍。相反,特异值对两个基于中位数的测量,四分位差 (*BDP* = 0.25) 和中位绝对离差 (*BDP* = 0.5) 的影响很小。与位置的 M 估计相似,特异值并未影响双权中位方差的表现,后者的 *BDP* 等于 0.5。

例 2.2:跨国视野下公众对报酬不平等的态度

下面我来看一个使用实际社会科学数据的例子。表 2.3 中的数据来自 Keakliem, Anderson 和 Heath(2005)对收入不平等与公众对报酬不平等的态度的关系的跨国研究。该数据集包括 20 世纪 90 年代测量的 48 个国家的信息。变量包括:

(1) *Secpay*。来自世界价值观调查(Inglehart et al., 2000)的一道题目的平均得分,这道题目询问了受访者对于报酬(secpay)不平等的态度。题目是这样表述的:"假设有两个秘书,她们年龄一样,实际上做的工作也一样。其中一个发现另一个挣的钱比自己多得多。只是报酬较高的那个做事更快,在工作上更有效率、更可靠。在您看来,这个秘书的报酬比另一个高是否公平?"受访者有两种回应选择:"公平"(编码为 0),或者"不公平"(编码为 1)。因此,较高的平均分反映公众更喜欢平等(也就是说,这个国家的多数受访者认为两个秘书的工资不同是不公平的)。这些国家的平均得分从 0.054 到 0.622,均值为 0.2。

(2) *Gini*。基尼系数,理论上的取值范围为 0(完全的收入平等,收入在所有国民之间平均分配)到 1(完全的不平等,一个人占有所有收入)。也就是说,取值越高表明收入越不平等。

(3) *Per Capita GDP/1000*。人均国民生产总值(千美元)。

(4) *Democracy*,民主。虚拟变量,"传统的民主国家(地区)"(也就是在收集数据时已经历过至少 10 年民主制度的国家)被编码为 1,"新的民主国家(地区)"被编码为 0。

用以建构这些测量指标的源数据的具体信息,可以参看 Weakliem et al.,(2005)。

表 2.3　48 个国家(地区)的公众态度及经济和政治变量

国家(地区)	Secpay	Gini	*Per Capita GDP/1000*	*Democracy*
亚美尼亚	0.061	44.4	2072	0
澳大利亚	0.179	31.7	22451	1
奥地利	0.112	23.1	23166	1
阿塞拜疆	0.070	36.0	2175	0
孟加拉国	0.057	28.3	1361	0
白俄罗斯	0.075	28.8	6319	0
比利时	0.302	27.2	23223	1
巴西	0.232	60.1	6625	0
英国	0.211	34.6	20336	1
保加利亚	0.164	30.8	4809	0
加拿大	0.176	28.3	23582	1
智利	0.361	56.5	8787	0
中国	0.131	41.5	3105	0
克罗地亚	0.092	29.0	6749	0
捷克共和国	0.557	26.6	12362	0
丹麦	0.248	21.7	24217	1
多米尼加共和国	0.089	50.5	4598	1
爱沙尼亚	0.054	35.4	7682	0
芬兰	0.354	22.6	20847	1
法国	0.231	32.7	21175	1
格鲁吉亚	0.086	37.1	3353	0
匈牙利	0.115	28.9	10232	0
印度	0.226	29.7	2077	1
爱尔兰	0.289	35.9	21482	1
意大利	0.226	34.6	20585	1
日本	0.284	24.9	23257	1
拉脱维亚	0.070	28.5	5728	0
立陶宛	0.096	33.6	6436	0
墨西哥	0.211	53.7	7704	0
摩尔多瓦	0.127	34.4	1947	0

（续表）

国家(地区)	Secpay	Gini	*Per Capita GDP/1000*	*Democracy*
荷兰	0.328	31.5	22176	1
挪威	0.441	24.2	26342	1
秘鲁	0.175	46.2	4282	1
葡萄牙	0.265	35.6	14701	1
罗马尼亚	0.133	28.2	5648	0
俄罗斯	0.076	48.0	6460	0
斯洛伐克	0.622	19.5	9699	0
斯洛文尼亚	0.108	29.2	14293	0
西班牙	0.286	32.5	16212	1
瑞典	0.401	25.0	20659	1
瑞士	0.149	36.1	25512	1
中国台湾地区	0.075	27.7	12090	0
土耳其	0.207	41.5	6422	0
乌克兰	0.085	47.3	3194	0
乌拉圭	0.273	42.3	8623	0
美国	0.148	36.9	29605	1
委内瑞拉	0.208	46.8	5808	1
联邦德国[a]	0.149	30.0	22169	1

注：a. 因为调查仅对联邦德国的受访者进行了调查，因此数据集使用了“联邦德国”这一概念。

令人感兴趣的是那些在调查时民主化不足10年的国家（或地区）($n = 26$)对报酬不平等的公众态度（下面简称为公众态度）的分布情况。由于公众态度变量将被作为后面回归分析的因变量，因此首先对它的分布进行探索是非常重要的，检查它是否具有一些可能带来问题的特征——如偏态或存在特异值。我们先来检查图2.2，它呈现的是公众态度变量的分布的核密度估计（kernel density estimation，也就是经过修匀的直方图）。除了在右端尽头有一个小鼓包外，分布的其他部分还是相当对称的。进一步的分析发现，两个国家——捷克共和国和斯洛伐克——的取值异常高。从表2.3中可以看到，这两个国家的取值分别为0.557和0.622，而其他国家的取值没有超过0.4的。它们的独特性可能与其共同的文化和历史遗产有很大关系。

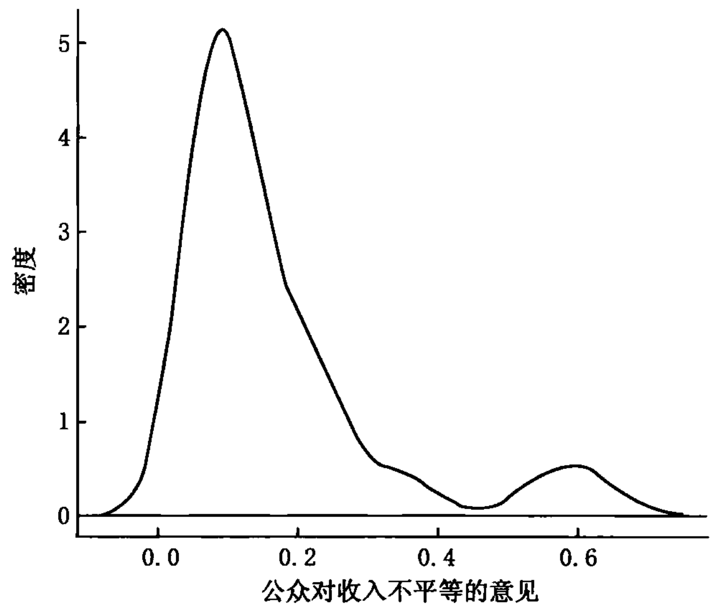


图 2.2　26 个新建民主国家(地区)的公众对报酬不平等的态度的分布情况

现在我们来看表 2.4,它探索了位置和尺度的各种测量在纳入和删除捷克共和国及斯洛伐克时的不同表现。首先来看均值,可以看到当特异值删除时它显著地变小了(从 0.167 变为 0.131)。类似的,当特异值被移除时基于均值的

表 2.4　新建民主国家(地区)的公众态度变量的位置及尺度测度

估计量	所有案例	捷克和斯洛伐克被移除
位置测度		
均值	0.167	0.131
α 截尾均值	0.123	0.114
中位数	0.112	0.102
M 估计	0.127	0.112
尺度测度		
标准差	0.145	0.078
平均离均差	0.102	0.060
平均离中差	0.081	0.056
四分位差	0.129	0.097
中位绝对离差	0.042	0.032
双权中位方差	0.005	0.004

尺度测度的缩小也很大(保留特异值时的标准差是删除特异值时的1.8倍)。相反,中位数及M估计的稳定性非常明显,它们的结果在特异值删除前后看起来几乎没有发生变化。类似的,两个数据集在中位绝对离差及M估计(双权中位离差)——两个具有高崩溃点的尺度测度——之间的差异要小得多。

在结束本章之时,有必要提醒的是:应该对就回归分析中的变量进行的单变量分布检查保持警惕。OLS回归估计的是在给定 x 下 y 的条件均值。因此,y 的特异值并不见得就是回归的特异值。反过来说,回归中的权势案例并不见得一定是 y 意义上的极端值。当然,这并不意味着我们应该忽视单变量分布情况。不对单变量的分布进行探索,可能会妨碍研究者发现数据的重要特征。但最好不要在变量之间的关系得到探索之前对异常观察案例进行处理。带着这样的想法,我们下面将转向线性回归OLS估计,具体讨论异常案例将如何影响它的估计结果,而这些案例又如何被甄别出来。在下文中,我们还会在稳健回归方法的背景下回到对尺度及位置测度的讨论。

第 3 章 | 稳健性、抗扰性与最小二乘回归

一般最小二乘回归

如果用 i 表示个体观察案例，j 表示变量，那么线性回归方程可以被写成：

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \qquad [3.1]$$

其中 y_i 是因变量或者说反应变量，x_{ij} 是用来预测 y 的自变量，β_j 是回归系数，ε_i 代表与各个 x 保持独立的随机项（也即误差项）。OLS 解将残差平方和最小化，其中残差是误差项的估计值。

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2 = \min \sum_{i=1}^{n} e_i^2 \qquad [3.2]$$

对方程 3.2 进行微分计算，得到：

$$\sum_{i=1}^{n} \left(y_i - \sum x_{ik}\beta_k\right)x_{ij} = \sum_{i=1}^{n} (e_i)x_{ij} \qquad [3.3]$$

所谓残差，期望值为 0，定义很简单，如下：

$$e_i = y_i - \hat{y}_i \qquad [3.4]$$

其中 $\hat{y}_i$ 为模型的预测值。

采用矩阵代数形式的话，线性回归模型就是：

$$y = X\beta + \varepsilon \qquad [3.5]$$

其中 $\underset{(n\times 1)}{y}$ 是反应变量的一个观察矢量，$\underset{(n\times k+1)}{x}$ 为包括所有案例每个解释变量取值的模型矩阵，$\underset{(k+1\times n)}{\beta}$ 为未知系数向量，$\underset{n\times 1}{\varepsilon}$ 为随机误差向量，即观察值 y 与其期望

值 $E(Y)$ 之差。

系数的最小二乘解为：

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad [3.6]$$

反应变量的拟合值或者说预测值，可由下面的式子求出：

$$\hat{y} = X(X^T X)^{-1} X^T y = Hy \qquad [3.7]$$

其中 H（"帽子矩阵"）是一个 $n \times n$ 的将 y 映射到预测值 $\hat{y}$ 上的对称矩阵。H 的对角线元素，也被称为预测值 h_i（hat value h_i），给出了每个观察案例的杠杆效应（leverage）。如果一个案例有异常的 x 值——也就是说它的取值远远高于或低于 x 的平均值，那它对回归平面的杠杆效应就高。后面我们将会看到，残差及帽子值提供了每个案例对回归系数的影响大小的重要信息。

线性模型的经典统计推论基于如下几个关于误差项的假定：

(1) 因变量 y 与解释变量之间的关系为线性关系。也就是说，在给定 x 值的情况下，ε 的期望值为 0，$E(\varepsilon_i) = 0$。

(2) 在所有 x 值下，误差项的方差恒定，$V(\varepsilon_i \mid x_i) = \sigma_\varepsilon^2$。通常又被叫做同方差性（homoscedasticity）假定。

(3) 误差项 ε_i 相互独立，$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$，其中 $i \neq j$。

根据高斯-马科夫定理（Gauss-Markov theorem），如果满足上述条件，那么 OLS 估计是总体回归系数的最佳线性无偏估计（best linear unbiased estimators, BLUE）（Draper and Smith 1998:136）。事实上，不管误差项的分布特征如何，$\hat{\beta}$ 都提供了方差最小的无偏估计。如果我们加上误差项服从正态分布的假定，也就是说 $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$，那么 $\hat{\beta}$ 能够提供 β 的最大似然估计，且其概率分布很容易推导出来。更具体地讲，如果误差项服从正态分布，那么系数也将服从正态分布，因为，它们是误差项的线性函数。不过，根据中心极限定理，对于 OLS 估计的标准误而言，正态性只有在 n 很小时才很重要。

回归系数的方差协方差矩阵是这样定义的：

$$V(\hat{\beta}) = (X^T X)^{-1} \sigma_\varepsilon^2 \qquad [3.8]$$

其中，误差项的方差 σ_ε^2 是通过残差方差 S_ε^2 来估计的。而系数的标准误是由等式 3.8 的对角线元素提供的。大的残差——能够反映特异值或重尾——会扩大 S_ε^2 估计，进而扩大估计值的标准误。特异值还能带来不恒定的误差方差，从而使 OLS 估计失去效率，因为它们给所有案例同样的权重，包括特异值，尽管后者含有较少的回归信息。正如我们后面即将看到的，模型矩阵 X 中的异常案例还对 OLS 估计的标准误有着重要的影响，效果是好是坏，取决于异常性的不同类型。（见 Cook and Weisberg，1999:161）。

异常案例对 OLS 估计及标准误的影响（Implications）

为了更好地理解异常案例是如何对回归估计产生影响的，首先来定义四个概念：单变量特异值（univariate outlier）、回归特异值（regression outlier）、杠杆效应（leverage）以及权势或影响（influence）（Cook and Weisberg，1982；又见 Rousseeuw and van Zomeran，1990；Fox，1991，1997）。所谓单变量特异值，是指远离某一变量分布中其他案例的观察值。尽管在进行回归分析之前的初步分析中对此类案例进行探测是审慎的做法，不过，它们并不一定会带来问题。换句话说，一个在 y 值或者 x 值上无条件异常（unconditionally unusual）的观察案例，并不必然是回归分析中的特异值。

回归特异值，有时又叫垂直特异值（vertical outlier）（Rousseeuw and van Zomeren，1990），处于数据主体的一般模式之外。更具体地讲，它是那种在 x 取值相同的情况下，y 取值非常不同的案例。回归特异值的典型特征就是残差很大。不过，残差大并不必然意味着该观察值影响了回归斜率的估计。同样，残差小并不一定意味着该观察案例服从数据主体部分的模式，至少对于来自 OLS 回归的残差是这样。当一个特异值的影响力（或权势）很大时，它可能将回归平面整个拉向自己，从而使得对应的残差很小。

一个案例如果 x 值异常，那它对回归平面就有杠杆效应。更详细地说，离 x 的平均值越远的观察案例（不管是在正方向，还是负方向），杠杆效应越大。不过，杠杆效应大不一定同时权势大。一个杠杆效应很大的观察案例可能恰好与

数据主体的模式一致。在这种情况下，该案例一点也不成为问题。后面很快就会有更具体的讨论。

如果将一个案例从回归中移除后，回归估计发生很大的变化，那么该案例就是有权势的。权势的大小取决于杠杆效应与 y 值的异常性的组合。也就是说，如果一个案例不仅有很高的杠杆效应，而且其 y 值在给定的 x 值下又很不正常，那它就能严重影响回归平面。在这种情况下，当回归平面试图把握这一观测案例时，截距和斜率都会受到影响。由于 OLS 估计建基于因变量的条件均值，会遇到均值本身会遇到的相同问题。仅仅一个案例就可以对系数估计产生影响。换句话说，OLS 估计的崩溃点 $BDP(T, Z) = 1/n \cong 0$（当 n 增大时，$1/n$ 趋近于 0），而它的影响函数与残差的大小成比例。哪怕只有一个问题案例存在，OLS 给出的估计也会被扭曲。

不同类型的异常案例还会对 OLS 估计的标准误产生影响。以简单回归模型为例，回归斜率的标准误建立在残差标准差 s_e 和 x 围绕其自身均值的变异量基础上：

$$S\hat{E}(\hat{\beta}) = \frac{se}{\sqrt{\sum (x_i - \bar{x})^2}} \qquad [3.9]$$

与重尾分布尾部的案例——包括特异值——相关的大残差会扩大 s_e 的取值，从而导致标准误比残差服从正态分布时大。具体地讲，杠杆效应小（即 x 值并不异常）的垂直特异值（即远离数据主体模式的观察案例）对上式的分母没有影响，但会扩大分子，进而使标准误变大。相反，带有杠杆效应的观察案例（即 x 取值在正方向上或负方向上远离其均值）将使分母变大，进而减小标准误。简单地讲，离其他数据较远但服从一般模式的观察案例能够提高 OLS 估计的精确性。只有在给定 x 值时 y 值异常的情况下，观察案例才会损害精确性。在这种情况下，一些稳健回归估计的标准误会更小。

图 3.1 展示了不同类型的偏差值对简单回归直线的影响。除了带标签的观察案例外，三个图中的数据完全一样。它们是人为设计出来的，因此当带标签的案例被排除时，x_i 和 y_i 之间有着很强的线性关系（见实线）。虚线表示的是受

污染数据(即带标签的数据纳入其中时)的回归。

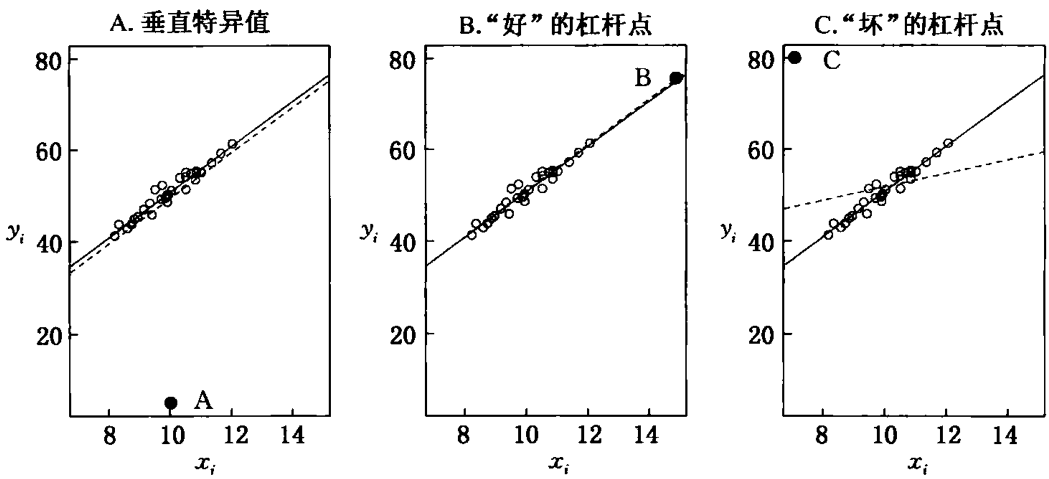


图 3.1 不同类型的偏差值及其对 OLS 估计的影响

注:虚线代表包括特异案例时的回归线;实线为不含特异案例的回归线。

首先来看图 A,观察案例 A 的 x 值并不异常(事实上,它等于 $\bar{x}$),但 y 值却高度异常。也就是说,这个案例是一个回归特异值,但没有很强的杠杆效应。因此,它对斜率的估计没有影响。不过,它确实对估计的精度产生了影响。由于残差如此之大,它会使标准误扩大。而且这种观察值会把截距拉向它,尽管只有当偏差非常极端时才会有严重的影响(在这里,它的影响很小)。在这种情况下,根据残差大小降低相应观察案例权重的稳健回归(如 M 估计)能够提供同样无偏但更加精确的估计。

在图 B 中,观察案例 B 不论是在 x 值上还是 y 值上都是单变量特异值,但它恰好落在回归直线上。也就是说,B 的杠杆效应很大,但并非回归特异值。尽管它对斜率没有影响,但它降低了估计的标准误,因为它拓宽了 x 的范围。因此,从 OLS 估计来讲,这个案例根本就没有问题。事实上,此时使用 OLS 之外的任何方法都是不明智的。

最后,图 C 中的案例 C,从 x 值上讲异常,在 x 值给定的情况下 y 也不正常。换句话说,C 的杠杆效应大,而且是个回归特异值,从而导致回归直线被拉向它。由于特异值影响而变大的 y 的变异性使得回归系数的精确性变小。如

何处理这种性质的特异值，需要良好的判断力，而且通常要进一步研究。在理想状态下，这种异常性是由于错误编码造成的，因此纠正编码即可。或者，如果有充足的理由，应该移除该案例，并将其作为特殊案例单独进行讨论。还有一种选择是使用某种形式的稳健回归——它们给出的结果通常与直接删除这些案例得到的结果基本相同。在目前这种特定的情况下，最好选择一种同时考虑了杠杆效应和残差的方法(如 GM 估计，将在第 4 章中讨论)。在后面我们即将看到，并非所有的稳健回归技术都有这种属性。

总的来讲，杠杆效应小的回归特异值尽管能影响截距的估计，但对回归斜率的影响很小。更重要的是，它会对模型的拟合以及估计的标准误造成负面影响，因为它的残差很大。一个杠杆效应很大但 y 值差异不大的案例——也就是说该案例与数据的主体模式保持一致——不会影响斜率估计，事实上，它的出现会增进模型的拟合度，让估计更加精确。只有当偏差与杠杆效应结合在一起时，斜率系数才会受到影响。

所有这些都表明，如果没有发现并处理好这些权势案例，将导致错误的结论。这些结论将建立在一个很差劲的模型之上，不管用来评价这个模型的各种标准测量——如 R^2 和系数标准误——看起来是否很好。也就是说，模型拟合度的标准测度量并非总能指示偏差案例对于回归系数的影响。这意味着使用图形法来评估案例的影响或权势非常重要。我们必须认真检查数据中的模式，从而对模型得出的估计抱有信心。当然，这是统计分析的一般原则，而非仅仅适用于回归分析。

例 3.1:26 个新建的民主国家(地区)的收入不平等与公众对报酬不平等的态度

没有其他办法比通过一个实际例子更能揭示 OLS 的局限性。继续使用上一章介绍的关于公众态度的数据，目标是用基尼系数来对公众态度进行预测。更准确地说，我们想研究一个国家(地区)的收入不平等水平是否会影响其公众对报酬不平等的态度。当下的例子只集中关注这 26 个新建民主国家及地区。

图 3.2 通过散点图对两个变量之间的关系进行了展示。图中还包括拟合这一数据的各种回归直线。其中，实线是利用全部 26 个案例信息来预测公众态度的 OLS 回归直线，虚线是排除了两个极端案例(捷克和斯洛伐克)后的 OLS 回归；点线来自于使用 M 估计的稳健回归，这种回归给予在最初 OLS 拟合中残差较大的观察案例更小的权重。(关于这种方法，更详细的内容将在第 4 章给出。)

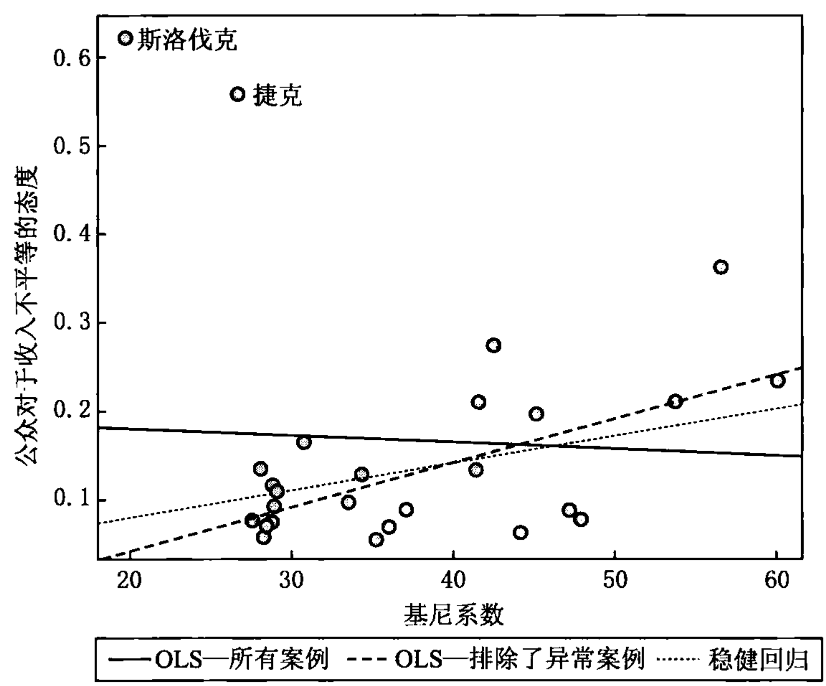


图 3.2 26 个新建民主国家(地区)的收入不平等与公众对于报酬不平等的态度

通过作图法对数据进行检查的重要性在这里非常明显，很容易就能看到：OLS 回归直线被拉向特异值。两个异常案例都具有很高的杠杆效应(也就是说，它们的 x 值，基尼系数，距离平均值 $\bar{x}$ 很远)，并且都是回归特异值(即：在给定的 x 值下它们都具有差异性很大的 y 值，公众态度)。也就是说，这两个案例并不符合数据主体的一般模式。我们还能看到，稳健回归比 OLS 回归更能反映数据中多数案例的模式，它对数据的拟合结果与剔除了特异值之后的 OLS 回归有一些类似。

上述回归的数字输出结果呈现在表 3.1 中。如果没对异常案例进行探测就报告最终的回归结果，我们很可能会选择 OLS 估计，并得出结论，认为收入不平等与公众对于报酬不平等的态度无关。一旦如此，我们就让两个案例——捷克和斯洛伐克——影响了整个结论。移除这两个案例之后的 OLS 回归表明，两个变量之间存在着相当强的关系。不仅基尼系数变成正的，而且在约定水平（$p<0.001$）上统计显著，回归模型对于数据的拟合也显著提高。在删除特异值之前，模型的确定系数接近于 0，删除后，确定系数增加到可观的 0.39。同样的，回归直线的标准误在特异值被删除后得到了显著的改善，从 0.149 下降到少于一半，约为 0.063。和图 3.2 反映的一样，类似于删除了特异值的 OLS 回归，稳健回归对于收入不平等的估计也是正的，尽管要小一点。影响在统计上仍然显著（$p = 0.025$），不过因为 n 很小，对表中提供的渐近标准误，阐释起来应该小心（在第 5 章将进一步讨论）。

表 3.1 以收入不平等来预测公众态度（新的民主国家和地区）

	OLS(所有案例)		OLS（排除捷克和斯洛伐克）		稳健回归(M 估计)	
	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)
截距	0.195	0.111	−0.059	0.053	0.016	0.056
基尼系数	−0.0008	0.0028	0.0050	0.0013	0.0031	0.0015
S_e	0.1485		0.0627			
R^2	0.0029		0.3887			
n	26		24		26	

探测 OLS 回归中的问题案例

上一章已经表明了在简单回归中探测和处理异常案例的重要性。现在我们来看一些用于探测多元回归分析特异值的传统方法。对于这些方法，绝大部分都会提供一些基本的描述。关于这里讨论的方法及其他方法的更多具体内容可以看库克和威斯伯格（Cook and Weisberg, 1982），鲁索和勒罗伊（Rousseeuw and Lerroy, 1987），查特吉和哈迪（Chatterjee and Hadi, 1988），还有福

克斯等的作品(Fox, 1991)。

我首先对上述跨国数据拟合了两个 OLS 回归模型,两个都是用基尼系数和人均 GDP 来解释公众态度。模型 1 包括所有 26 个新的民主国家(地区)。如表 3.2 所示,基尼系数对于公众态度的影响的统计检验不显著 ($p=0.40$),但人均 GDP 的影响统计显著 ($p=0.018$)。此前的分析表明,两个案例——捷克和斯洛伐克——在基尼系数取值很低的情况下有着异常高的公众态度得分。排除了这两个案例的模型 2 得出了不同的实质性结论。人均 GDP 不再显著 ($p=0.051$),而且大小只有包含两个特异值时的 1/3。相反,基尼系数的影响增长了 7 倍,而且变得统计显著了 ($p=0.002$)。同样重要的是,包含两个异常案例的回归的标准误是没有包含时的两倍以上,而确定系数 R^2 不到后者的一半。因此,很清楚的是,这两个案例对估计造成了严重的问题。当然,在进行模型诊断之前,我们事先通常并不知道异常案例存在。下面我们对使用了全部 26 个观察案例(包括两个特异值)的回归进行诊断。

表 3.2 用收入不平等预测公众态度的不同 OLS 回归来自新的民主国家(地区)的基尼系数和人均 GDP 的解释

	OLS(所有案例)		OLS(排除捷克和斯洛伐克)	
	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)
截距	0.028	0.128	−0.107	0.058
基尼系数	0.00074	0.0028	0.00527	0.0013
人均 GDP(千美元)	0.0175	0.0079	0.0063	0.0037
S_e	0.138		0.0602	
R^2	0.175		0.4622	
n	26		24	

探测杠杆效应:帽子值(Hat value)

我们先来探测杠杆效应高(即 x 取值比较异常)的案例。最常用的衡量杠杆效应的测量是前面已经讨论过的(见方程 3.7)帽子值 h_i,它是将特定 y_i 转化为对应预测值 $\hat{y}_i$ 的权重。如果 h_{ij} 很大,那么第 i 个观察案例对第 j 个预测值有

很大的影响：

$$\hat{y}_i = h_{1j}y_1 + h_{2j}y_2 + \cdots + h_{nj}y_n = \sum_{j=1}^{n} h_{ij}y_i \qquad [3.10]$$

单个帽子值 h_i 衡量的是特定观察案例 y_i 对所有预测值的潜在杠杆效应。帽子值的范围处于 $1/n$ 和 1 之间，平均值为 $\bar{h} = (k+1)n$。值得注意的是，在最小二乘回归中，y 的取值与帽子值的计算无关。[9] 帽子值只考虑了一个特定的 x_i 值离它的均值 $\bar{x}$ 有多远。如在简单回归中，

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \qquad [3.11]$$

帽子值在多元回归中的功能相同，不过 h_i 现在测量的不是相对于单个 x 的均值的距离，而是相对于多个 x 的中心的距离——即所有 x 的均值相交的地方。换句话说，多元回归中的帽子值考虑了各个 x 之间的相关及变化结构(correlation and variational structure)(见 Cook and Weisberg, 1999: 161—163)。帽子值只能指示杠杆效应——它们不能告诉我们在给定的 x 取值下对应的 y 值是否异常。尽管杠杆效应高的案例有时会有很大的残差，但并不必然如此。事实上，正因为它们将回归直线拉向自己，杠杆效应高的点残差可能很小，

$$V(e_i) = \sigma_\epsilon^2(1 - h_i)$$

图 3.3 标示出了模型 2 中与案例号对应的帽子值(即所谓的带标号的帽子值图)。尽管这里并没有针对杠杆效应的正式检验，经验之一是，超过帽子值平均值两倍的帽子值应该被关注。[10] 图中的虚线即代表对应取值。有三个案例与其他数据差别较大：巴西、智利和斯洛文尼亚。尽管这些案例的杠杆效应很高，但没有进一步的分析，还不能说它们偏离了数据主体的模式。回想一下，如果它们与数据中其他案例一致的话，这些案例就是“好的”杠杆案例，因为它们帮助降低了估计的标准误。因此，在进一步探测回归特异值时，我们会留意这些案例，但不用对它们有所担心。

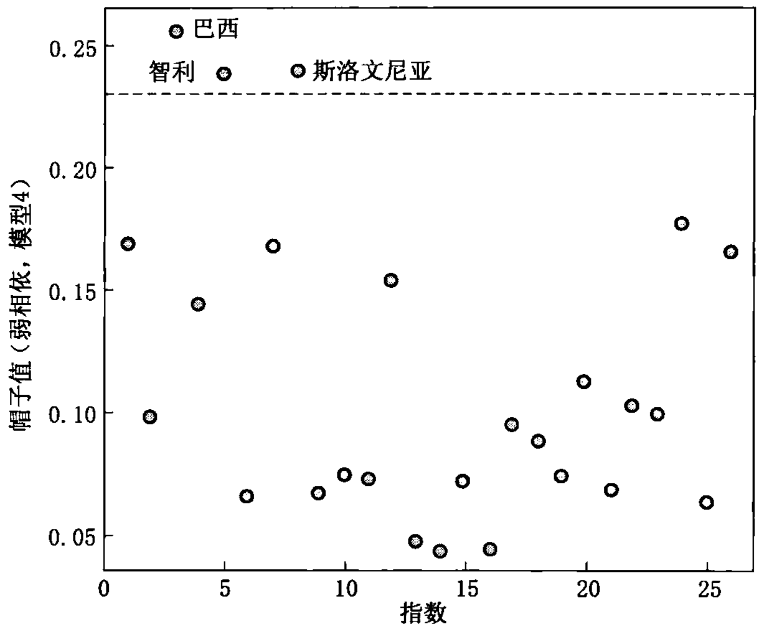


图 3.3 模型 1 的帽子值

探测回归特异值：学生化残差及邦弗鲁尼校正（Studentized Residuals and Bonferroni Adjustment）

乍一想，探测回归特异值最简单的办法是计算标准化的残差 e'_i，然后将 $|e'_i|>2$ 的案例（也就说，大于平均残差两个标准差的残差）作为显著的特异值。但是，标准化残差值在进行统计推论时存在麻烦，因为等式的分子和分母相互并不独立，因而并不服从 t 分布。如下所示，我们关注的残差 e_i 不仅出现在分子里，还出现在分母中的残差标准差 s_e 的计算里：

$$e'_i=\frac{e_i}{s_e\sqrt{1-h_i}} \qquad [3.12]$$

解决办法是计算学生化残差（studentized residuals），e'_i，它将我们关注的案例从残差标准差的计算中排除掉：

$$e^*_i=\frac{e_i}{s_{e(-i)}\sqrt{1-h_i}} \qquad [3.13]$$

它满足自由度为 $n-k-2$ 的 t 分布。

将学生化残差与自由度相同的理论 t 分布量进行比较的分位对比图(quantile comparison plot)对探测特异值非常有用，尤其是在拟合回归模型之前，并没有任何特定的观察案例被怀疑为特异值的情况下，事实上，通常都是这种情况。图 3.4 是来自模型 1 的残差的分位对比图。围绕这些观察值的 95%的置信带(confidence envelop)是使用自助法得到的，这一主题将在第 5 章得到讨论。有两个案例偏离了置信带：捷克和斯洛伐克。回想一下可知，这两个案例在前面的诊断中并没有被认为带有极端的帽子值。

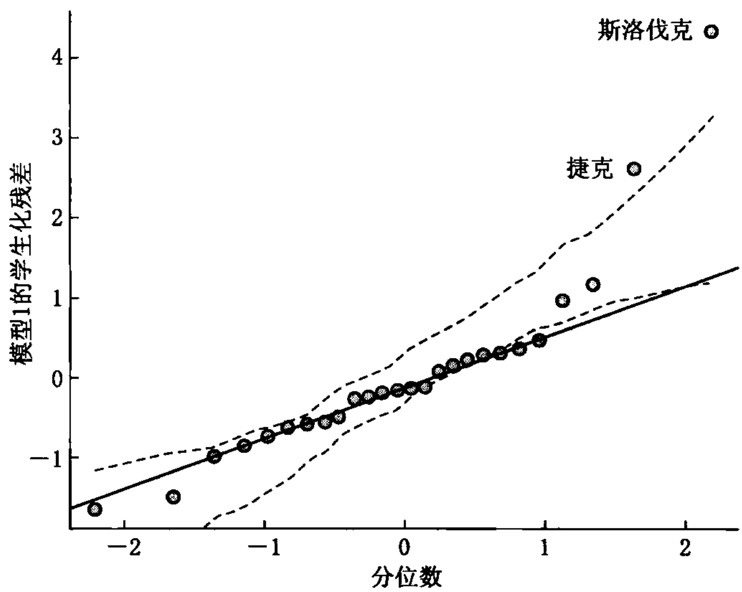


图 3.4　来自模型 1 的学生化残差分位对比图

我们可以正式地检验一个具体的案例是否特异值，尽管标准 p 值不能信任。如果我们有意地检验最极端的残差——事实上，经常这样做——而不是随机选择一个观察案例，检验将偏向于统计显著，因为即使残差满足正态分布，仅仅因为随机的原因就有 5%的学生化残差会在统计检验上显著 ($\alpha = 0.05$)。对最大特异值的双边 t 检验的 p 值进行 Bonferroni 校正能够补救这一问题。Bonferroni p 值 $= np'$，其中 p'是没有被调整过的自由度为 n-k-2 的 t 检验的 p 值(Fox，1997:274)。斯洛伐克——残差最大的观察案例——的 t 统计量等于 4.31，对应自由度为 22，p 值 $= 0.00027$。经过 Bonferroni 校正，p 值为 26 ×

0.00027 = 0.0072，表明斯洛伐克是一个显著的特异值。尽管这里我们发现了偏差很大的案例，但必须记住，异常案例并不总带有很大的残差。

探测权势值：DFBETAs、Cook's D 和偏回归图（partial regression plot）

评估权势值最直接的办法是看特异值移除后回归系数发生了多大变化。我们可以使用被贝尔西、库恩和韦尔斯（Belsley, Kuhn and Welsch, 1980）称作 DFBETAs（或 D_{ij}）的概念，它们的定义非常简单：

$$D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}, \text{对于 } i = 1, \cdots, n;\ j = 0, 1, \cdots, k$$

其中 $\hat{\beta}_j$ 是使用所有数据时的回归系数，$\hat{\beta}_{j(-i)}$ 是第 i 个案例被移除时的系数。每个案例对于每个系数都有一个 $DFBETA_{ij}$。对于 $DFBETA_{ij}$ 没有正式的显著性检验，常用的拇指规则（rule of thumb）界线是 $|D_{ij}| \geq 2/\sqrt{n}$。将 $DFBETA_i$ 值对案例编号作图①很有用，然后从中寻找取值相对较大的。

回到模型 1，人均 GDP 及基尼系数两个变量对应的 $DFBETA_i$ 值标记图（index plot）呈现在图 3.5 中。点线代表拇指规则界线，$|D_{ij}| \geq 2/\sqrt{n}$。我们已经发现捷克和斯洛伐克是回归特异值。现在我们看到它们对回归估计还有异常高的影响。这种影响同时涉及两个解释变量。斯洛伐克明显将人均 GDP 的系数往正方向拉，将基尼系数的影响往负方向拉。尽管与斯洛伐克相比，捷克共和国对人均 GDP 的影响较大而对基尼系数的影响较弱，但两者的一般模式是相同的。另外三个潜在的问题案例也被发现：智利将基尼系数往正方向拉；中国台湾地区和斯洛文尼亚将人均 GDP 的系数往负方向拉。

DFBETA 对于理解一个案例对哪个自变量存在权势非常有用，但是由于需要在每个变量上给每个案例一个单独的测量，当维度增多时（即自变量增加时），这种测量显得非常繁杂。Cook's 距离，也被称作 Cook's D，通过提供一个单一指标，测量每个观察案例对于回归平面的总体影响，解决了这一麻烦（Cook, 1977）。与 Cook'D 相关的其他方法还有 DFFIT（Belsley et al., 1980）、

---

① 即分别以案例编码和 $DFBETA_i$ 作为横、纵坐标做散点图。——译者注

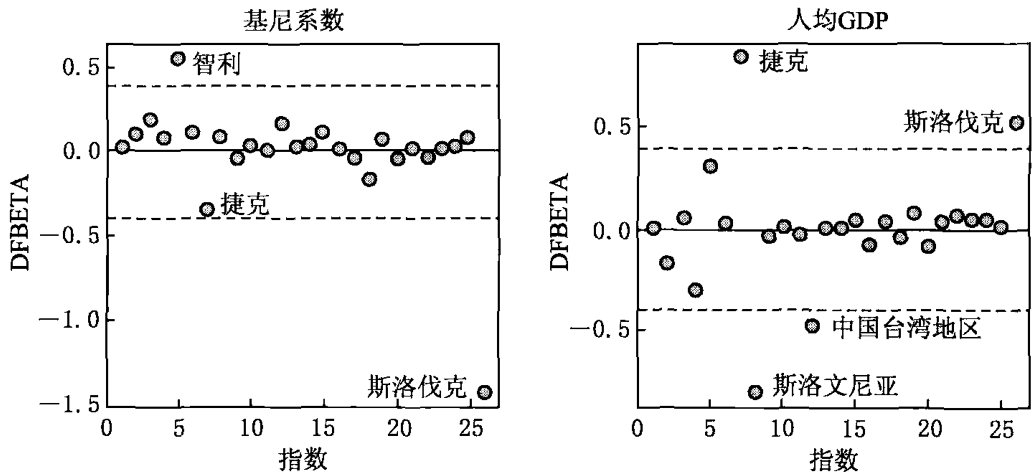


图 3.5 模型 1 中人均 GDP 及基尼系数的回归系数的 DFBETA 标记图

艾金森的调整 Cook's 统计量(Atkinson, 1985)。这些方法一般会给出类似的结果(见 Draper and Smith, 1998:214),因此我们主要关注最常用的 Cook's D。

某个观察案例的 Cook's D 为:

$$D_i = \frac{e_i'^2}{k+1} \times \frac{h_i}{1-h_i} \qquad [3.14]$$

其中 k 是自变量的个数。等式的前半部分包括该案例的标准化残差,e_i',从而测量了差异度(discrepancy)。等式的后半部分包括帽子值,h_i,因而测量了杠杆效应。尽管 Cook's D 没有正式的统计显著性检验,但有大致的标准分界线。Cook 和 Weisberg(1999:358)建议要认真探测 $D_i > 0.5$ 的观察案例。基于 Chatterjee 和 Hadi(1988)提供的与此相关的 DFFIT 测量的分界线,Fox(1997:281)建议了一个取决于样本规模 n 和自变量数目 k 的分界线 $D_i > \frac{4}{n-k-1}$。这些分界线都是有用的,但它们并非总能成功地区分出问题案例,因此,没有什么可以替代对于相对差异度的认真检查。

Cook's D 标记图能够帮助我们确定观察案例对回归估计的总体影响的相对大小(Cook and Weisberg, 1999:358)。另外一种选择是绘制一个由福克斯(Fox, 1991:37—38)创制的"权势图"[11]。权势值图的一个用处在于,它能够显示出决定权势大小的差异度及杠杆效应的相对权重。权势图是通过将学生

化残差 e_i^* 作为纵轴，帽子值 h_i 作为横轴作图得到的。每一个案例由一个空心圆圈代表，圆圈的面积与 Cook's D 成比例。也就是说，圈越大，对应案例对回归平面的影响越大。

图 3.6 给出了模型 1 的 Cook's D 标记图和权势图。两个图中的异常案例

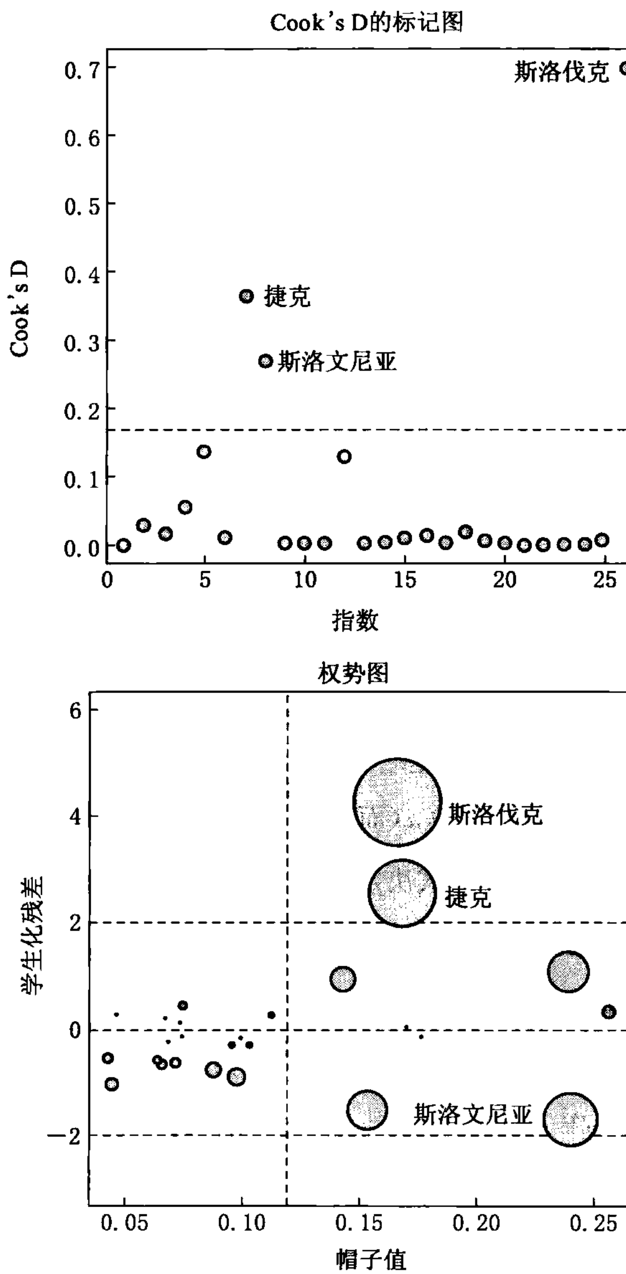


图 3.6　对模型 1 中各案例的权势进行评估

都加了标签。斯洛伐克和捷克的高权势通过权势图中代表它们的较大圆圈展现得非常清楚。而且我们可以非常清楚地看到，这两个案例的高权势，都是因为较大的差异度和杠杆效应相结合而导致的。我们还得注意，斯洛文尼亚对回归估计的影响同样比其他国家大得多。

当权势案例相对较少时，Cook's D 对它们的甄别非常成功，但它有可能不能发现联合性权势案例(jointly influential observation)，尤其是当这样的案例为数不少时。在这种情况下，有可能这些案例中的任何一个本身并没有高度异常的权势，因此也不会有很大的 Cook's D。如果联合性权势案例数量很少的话，Cook's D 仍然可以被用来甄别它们，方法是逐步删除这些权势案例并不断更新模型，且每次都检查一下 Cook's D。但这种方法在需要试探的次数很多时变得不切实际。因此，偏回归图，又叫新增变量图(added variable plots)，要有用得多，至少在甄别对单个回归系数影响很大的观察案例上(Cook and Weisberg, 1999:360)[12]。

偏回归图和将所有案例标绘在图上的简单散点图很相似。不同的是，在偏回归图中观察案例呈现出来的模式表示的是 y 和 x 之间的偏关系(partial relationship)，而不是边际关系(marginal relationship)。也就是说，偏回归图画的是控制其他自变量不变的情况下一个变量的效果。令 $y_i^{(1)}$ 表示除 x_1 外所有其他 x 对 y 的最小二乘回归的残差，$y_i = a^{(1)} + \hat{\beta}_2^{(1)} x_{i2} + \cdots + \hat{\beta}_k^{(1)} x_{ik} + y_i^{(1)}$。类似的，令 $x_i^{(1)}$ 表示来自 x_1 对所有其他 x(而不是 y)的回归的残差，$x_{i1} = c^{(1)} + d_2^{(1)} x_{i2} + \cdots + d_k^{(1)} x_{ik} + x_i^{(1)}$。这两个等式决定着残差 $y_i^{(1)}$ 和 $x_i^{(1)}$ 是 y 和 x_1 在 x_2，…，x_k 的线性效应被移除之后剩下的部分。如果用来评价杠杆效应，残差 $y_i^{(1)}$ 和 $x_i^{(1)}$ 有 3 个极方便的属性：(1) $y_i^{(1)}$ 对 $x_i^{(1)}$ 回归的斜率就是多元回归中的最小二乘斜率 $\hat{\beta}_1$(也就是说，与偏回归斜率相等)；(2) $y_i^{(1)}$ 对 $x_i^{(1)}$ 回归的残差恰好等于最初多元回归的残差，$y_i^{(1)} = \hat{\beta}_1 x_1^{(1)} + e_i$；(3) $x_i^{(1)}$ 的方差是控制所有其他 x 不变时 x_1 的条件方差。正是因为这些属性，以 $y_i^{(1)}$ 对 $x_i^{(1)}$ 作图就能同时反映案例对 $\hat{\beta}_1$ 的单独权势和联合权势。

模型 1 的偏回归图在图 3.7 中。可以看到，偏回归图更清楚地说明了捷克

和斯洛伐克的权势，反映出它们一起将基尼系数的斜率拉向负方向，同时将人均 GDP 的斜率拉向正方向。换句话说，如果组合在一块，这两个案例的影响比它们各自单独的 Cook's D 显示的要大得多。

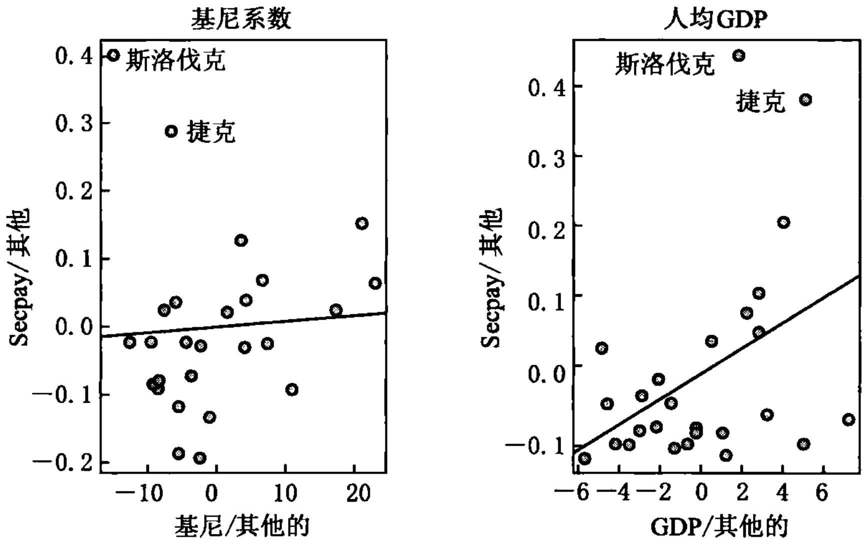


图 3.7　模型 1 的偏回归图

处理权势案例的一些策略

一旦在 OLS 回归中发现问题案例，就有几种选择：(1)审查一下这种偏差是不是通过重新编码或移除这些案例就可以解决的模型失效(model failure)的征兆；(2)转换一个变量(或多个变量)以矫正这个问题，尤其是碰到偏态分布时；(3)如果存在很多联合性权势案例，我们可以通过在模型中增加新的因素——要么是新变量，要么是已有自变量间的交互作用——来解决这些观察案例展现出来的异常模式；或者(4)使用一种更稳健、更能抵抗特异值干扰的方法。下面我们就将注意力转向各种稳健回归技术。

第 4 章 | 线性模型的稳健回归

现在我们来探讨各种稳健回归技术——包括那些有时被称为耐抗回归的技术，我们使用的是一种进化论的讨论方式，将分别解释各种新方法是如何来应对已有方法的局限性的。几种类型的回归将被讨论到：L 估计（基于序次统计量的线性组合），R 估计（基于残差的秩序），M 估计（通过考虑残差的大小由位置的 M 估计扩展而来），GM 估计（或者叫广义 M 估计，通过给予高权势点和大残差点较小的权重对 M 估计所做的扩展），S 估计（这种估计将残差尺度的稳健 M 估计最小化），MM 估计（同时基于 M 估计和 S 估计，以同时获得高的崩溃点和高的渐近效率）。其中有些方法已经被认为过时了，但我们仍会提供一般的描述，因为稳健回归技术的最新发展建立在它们之上。本章结尾时将讨论，稳健回归何以能够成为甄别问题案例的诊断方法。

L 估计量（L-estimators）

任何由顺序统计量的线性组合计算而来的统计量都可以被归类为 L 估计量。第一个 L 估计程序比 OLS 略微稳定一些，是最小绝对值（LAV）回归。最小绝对值回归又被称为 L_1 型回归[13]，因为它将 L_1 标准（L_1-norm，即绝对离差之和）最小化，这种方法是最简单也是最早用来限制案例权势的稳健回归，比 OLS 回归还要早产生 50 年（Wilcox，2005：451）。最小二乘回归也满足 L 估计的这一定义，因此有时也被称为 L_2 型回归，表示 L_2 标准（L_2-norm，离差平方和）被最小化。其他知名度比较高的 L 估计还有最小二乘中位数估计和最小截尾二乘估计。[14]

最小绝对值回归(Least Absolute Values Regression)

最小绝对值(LAV)回归对带异常 y 值的观察案例有很强的抗扰力。它是通过将残差绝对值之和最小化来求解的。

$$\min\sum_{i=1}^{n}|e_i| = \min\sum_{i=1}^{n}\left|y_i - \sum x_{ij}\beta_j\right| \qquad [4.1]$$

LAV 可以被看作是更一般的分位数回归的特例。在这种情况下,需要最小化的目标函数可以被写成:

$$\sum_{i=1}^{n}\rho_\alpha(e_i) \qquad [4.2]$$

其中:

$$\rho_\alpha(e_i) = \begin{cases} \alpha e_i & \text{如果 } e_i \geqslant 0 \\ (\alpha-1)e_i & \text{如果 } e_i < 0 \end{cases} \qquad [4.3]$$

而 α 是需要估计的分位数。有关分位数回归的一般应用,可以看凯恩克和巴萨特等的作品(Koenker and Bassett, 1978; Koenker and d'Orey, 1994; Koenker, 2005)。而针对社会科学的应用,可以看郝玲馨与奈曼的合著(Hao and Naiman, 2007)[15]。

尽管 LAV 不像 OLS 那么容易受异常 y 值影响,但它无法处理杠杆效应(Mosteller and Tukey, 1977:366),因此其崩溃点 $BDP = 0$。另外 LAV 的估计效率相对较低。与平均值的情况一样,在 y 服从正态分布 $y \sim N(\mu, \sigma^2)$ 的假定下,OLS 回归中 y 的抽样方差是 σ^2/n;但对于 LAV 回归来说,是前者的 $\pi/2 = 1.57$ 倍,等于 $\pi\sigma^2/2n$(也就是说,只有 OLS 估计效率的 64%)。LAV 回归法不仅崩溃点低而且效率低,这使得它不如接下来要讨论的稳健回归方法有吸引力。

最小二乘中位数回归(Least Median of Squares Regression)

最小二乘中位数法(LMS)最先由鲁索(Rousseeuw 1984)提出[16],它把

OLS 回归中的残差平方的总和替换为残差平方的中位数。估计结果是这样解出来的：

$$\min M\left(y_i - \sum xi_j\beta j\right)^2 = \min M(e_i^2) \qquad [4.4]$$

其中 M 表示中位数。其想法是通过将总和替换为更稳健的中位数，使得最终的估计量能够更好地抵抗特异值。尽管这一目标达到了(它的崩溃点 $BDP = 0.5$)，但 LMS 估计有非常重要的不足之处，以至于限制了它的广泛应用。它最好的相对效率也只有 37%(参见 Rousseeuw and Croux，1993)，另外，由于收敛速率只有 $n^{-1/3}$，它没有一个定义清晰的影响函数(Rousseeuw，1984)。尽管有这些不足，后面我们仍将看到，LMS 估计在效率高得多的 MM 估计的运算中扮演着非常重要的角色，它为后者提供了残差的初始估计。

最小截尾二乘回归(Least Trimmed Squares Regression)

另一种由鲁索发展出来的方法是最小截尾二乘回归(LTS)。LTS 回归由截尾均值扩展而来，通过最小化截尾残差平方和来求解。求解方式如下：

$$\min \sum_{i=1}^{q} e_{(i)}^2 \qquad [4.5]$$

其中 $q = [n(1-a)+1]$ 是估计量运算过程中包含的观察案例数，α 是截尾了的比例。使用 $q = (n/2)+1$，可以确保估计量的崩溃点 $BDP = 0.5$。尽管具有极高的抗异常值干扰的能力，不过 LTS 回归的相对效率方面非常差，只有大概 8%左右(见 Stromberg，Hossjer and Hawkins，2000)。它的效率如此之低，甚至不适合作为单独的估计量。但它在其他估计量的计算中有着重要作用。例如，由寇克里和赫特曼斯伯格(Coakley and Hettmansperger，1993)提出的 GM 估计就是使用 LTS 来取得残差的初始估计的。LTS 残差还能被有效地用于特异值诊断作图，后面将会讨论到。

R 估计量(R-Estimators)

R 估计量最先由贾克尔(Jackel，1972)提出，它们与基于序次化残差线性组

合(即基于残差的秩)的离散程度度量紧密相关。令 R_i 代表残差的秩 e_i，R 估计最小化的是秩化残差(ranked residuals)某种得分之和：

$$\min \sum_{i=1}^{n} a_n(R_i)e_i \qquad [4.6]$$

其中 $a_n(i)$ 是满足如下条件的单调得分函数：

$$\sum_{i=1}^{n} a_n(i) = 0 \qquad [4.7]$$

前人已经提出多种可能的得分函数。其中，最简单，可能也是用得最多的是威尔考克森得分(Wilcoxon score)，它直接寻找观察案例相对于中位数的秩：

$$a_n(i) = i - \left(\frac{n+1}{2}\right) \qquad [4.8]$$

中位得分是威尔考克森得分简单调整后的结果：

$$a_n(i) = \sin\left[i - \left(\frac{n+1}{2}\right)\right] \qquad [4.9]$$

范·德·瓦尔登得分(Van der Waerden score)以正态概率密度函数的反函数 Φ^{-1} 对秩进行了修正：

$$a_n(i) = \Phi^{-1}\left(\frac{i}{n+1}\right) \qquad [4.10]$$

最后，有限正态得分(Bounded Normal Scores)根据一个常数 c 对范·德·瓦尔登得分进行了限定和修正。

$$a_n(i) = \min\left\{c,\ \max\left[\Phi^{-1}\left(\frac{i}{n+1}\right),\ -c\right]\right\} \qquad [4.11]$$

R 估计量相对于其他估计量(如 M 估计及其扩展而来的各种估计量)的一个优点，在于它们具有尺度同变性。不过，它们也有一些不足之处。问题一是，这些得分函数哪个最优并不清楚。另一个问题是，对于截距而言，它们的目标函数不具可变性。如果不需要截距的话，也就无所谓了——事实上就没有估计截距。即使有人需要截距，也可以在拟合了模型之后通过残差的中位数手工计

算得到，因此这一点不足是可以克服的。更严重的问题在于多数 R 估计量的崩溃点 $BDP = 0$。一个特例是纳兰霍和赫特曼斯伯格（Naranjo and Hettmensperger，1994）的有限影响 R 估计量，它在高斯—马科夫假定满足的情况下效率也很高（90%—95%）。不过，即使对于这个估计量，它的崩溃点也从来没有达到超出 0.20 的水平。因此，我们先把 R 估计量放在一边，接下来讨论其他更稳健的估计量。（有关 R 估计更具体的内容，可以看休伯等的著作[Huber，2004；Davis and Mckean，1993；Mckean and Vidmar，1994]）。

M 估计量（M-Estimators）

回归的 M 估计首先由休伯（Huber，1964，1973，2004）提出，是位置的 M 估计的一种相对直接的扩展。它是首先对最小二乘估计量的效率与 LAV 估计量的耐抗性进行折中整合尝试的代表之一，后二者可以被看作是 M 估计的特例。简单地讲，M 估计量将残差的某种函数最小化。和位置的 M 估计一样，估计量的稳健性取决于权重函数的选择。

如果线性、方差齐性及独立误差假定成立，那么 β 的最大似然估计结果就等于使用最小化平方和函数求得的 OLS 估计：

$$\min \sum_{i=1}^{n} \left(y_i - \sum x_{ij}\beta_j\right)^2 = \min \sum_{i=1}^{n} (e_i)^2 \qquad [4.12]$$

与位置的 M 估计一脉相承，稳健回归 M 估计最小化的不是残差平方和，而是另外一个递增速率较低的残差函数之和：

$$\min \sum_{i=1}^{n} \rho\left(y_i - \sum x_{ij}\beta_j\right) = \min \sum_{i=1}^{n} \rho(e_i) \qquad [4.13]$$

这个函数的解不具备尺度同变性，因此残差必须经过标准化，而标准正是它们的尺度 $\hat{\sigma}_e$ 的某种稳健估计，这一估计是同时完成的。和在位置的 M 估计中一样，中位绝对离差（MAD）经常被使用。计算方程 4.13 的导数，解出来之后即为得分函数：

$$\sum_{i=1}^{n}\Psi(y_i-\sum x_{ij}\beta_j/\hat{\sigma})x_{ik}=\sum_{i=1}^{n}\Psi(e_i/\hat{\sigma}_e)x_i=0 \qquad [4.14]$$

且 $\Psi=\rho'$。这样，上式就变成了 $k+1$ 个方程形成的方程组，其中 Ψ 被替换为适当的随着残差增长而降低的权重：

$$\sum_{i=1}^{n}w_i(e_i/\hat{\sigma}_e)x_i=0 \qquad [4.15]$$

迭代再加权最小二乘法(Iteratively Reweighted least squares)

求解回归的 M 估计必须使用迭代程序。仅仅一步是没法求出结果的，因为残差在建立模型之前根本就不可能知道，而估计结果在不知道残差之前也求不出来。因此，迭代再加权最小二乘法(IRLS)被用来解决这一问题[17]：

(1) 设定迭代游标为 $I=0$，此时对数据估计一个 OLS 线性回归，找到回归系数的初始估计 $\hat{\beta}^{(0)}$。

(2) 从初始的 OLS 回归中算出残差 $e_i^{(0)}$，并用来计算初始权重。

(3) 选择一个权数函数，并将之用于初始的 OLS 残差，产生出预备(*preliminary*)权数，$w(e_i^{(0)})$。

(4) 第一次迭代，$I=1$，用加权最小二乘法(Weighted least squares, WLS)最小化 $\sum w_i^{(1)}e_i^2$ 并且得到 $\hat{\beta}^{(1)}$。以矩阵的形式表示，如果 W 为代表个体权重的 $n\times n$ 对角矩阵，其解为：

$$\hat{\beta}^{(1)}=(X^TWX)^{-1}X^TWy \qquad [4.16]$$

(5) 程序将继续使用初始的 WLS 回归得到的残差计算新的权重，$w_i^{(2)}$。

(6) 新权重 $w_i^{(2)}$ 将用在下一次 WLS 迭代中，$I=2$，估计出 $\hat{\beta}^{(2)}$。

(7) 第 4—6 步将被不断重复，直到 $\hat{\beta}$ 稳定在一次迭代结果上。

更一般地讲，q 次迭代每一次的解都是 $\hat{\beta}^{(I)}=(X^TW_qX)^{-1}X^TW_qy$，其中 $\underset{(n\times n)}{W_q}=\mathrm{diag}\{w_i^{I-1}\}$。迭代将一直持续，直到 $\hat{\beta}^{(I)}-\hat{\beta}^{(I-1)}\cong 0$。通常来讲，当估计结果的变

化量不超过上一次迭代的0.01%时，解被认为得到收敛。在第6章稳健广义线性模型部分，我们将更详细地讨论迭代再加权最小二乘法IRLS。

根据M估计的定义，它对重尾误差分布和不定误差方差——因而也对y特异值——具有耐抗性，不过它们仍然潜在地假定模型矩阵X不存在测量误差。但是在高斯—马科夫假定下，M估计的效率达到OLS估计的95%。[18]另外，虽然M估计在面对回归特异值(即在给定各个x值情况下异常的y值)时的稳健性及抗扰能力上比OLS回归有所改进，但和LAV估计量一样，它们不是完全不受异常案例的影响，因为它们没有考虑杠杆效应。回想一下可以知道，位置的M估计高度稳健，拥有有界影响函数并且崩溃点取值$BDP=0.5$。回归的M估计分享了关于y的那些属性，而没有继承关于x的那些属性，结果其崩溃点$BDP=0$。也就是说，在某些情况下它们的表现并不比OLS好(参见Rousseeuw and Leroy, 1987)。后面我们将会看到，因为它们在计算其他更稳健的估计量中的角色，这些估计量仍然非常重要。

GM估计量 (GM-estimators)

M估计由于未能处理杠杆效应，因而不具备有界的影响函数(Hampel et al., 1986)。为了应对这一问题，有界影响广义M估计(bounded influence Generalized M-estimator)被提了出来。其目标是产生出能够同时考虑垂直特异值和杠杆效应的权数。其中特异值经由标准的M估计处理，而杠杆点则按照通常的方式根据各自的帽子值赋予较轻的权数。GM族估计量的一般形式是这样定义的：

$$\sum_{i=1}^{n} w_i(x_i)\Psi\left\{\frac{e_i}{v(x_i)\hat{\sigma}_e}\right\}x_i=0 \qquad [4.17]$$

其中Ψ为得分函数(和在M估计中一样，通常被叫做Huber或双权函数)，权数w_i和v_i最开始取决于对数据拟合的初始OLS回归的模型矩阵X，然后在迭代中得到更新。

第一个GM估计由马娄斯(Mallows)提出(参见Krasker and Welsch, 1982)，它只包括w_i权数——也就是说公式4.17中$v_i(x_i)=1$。权数w_i由帽子

值计算而来。因为帽子值的取值范围为0到1,所以权数 $w_i = \sqrt{1-h_i}$ 保证了杠杆效应高的观察案例得到的权重比杠杆效应小的案例小(即如果 $h_i > h_j$,则 $w_i < w_j$)①。这一处理策略初看起来似乎很有道理,但实际上是有问题的,因为即使与数据主体的模式一致的"好的"杠杆点的权重也被削减了,从而造成了效率的损失。

史威普(Schweppe)提出的解决办法(在汉斯金等人的著作中有介绍,参见Handschin et al.,1975)是根据残差 e_i 的大小来对杠杆效应权数进行调整。为了得到这一结果,权数 w_i 的定义与马娄斯的一样 $w_i = \sqrt{1-h_i}$,但现在 $v_i(x_i) = w_i$(见Chave and Thomson,2003)。尽管史威普估计量的崩溃点比没有考虑杠杆效应的常规M估计的崩溃点要高,不过马荣娜、布托斯及约哈依(Maronna, Butos and Yohai,1979)表明它的崩溃点从来没有高过 $1/(p+1)$,其中 p 为模型要估计的参数的个数。换句话说,随着维数(dimentionality)增加,模型的崩溃点将向 $BDP = 0$ 靠近。这尤其会带来麻烦,因为随着模型中的变量增加,对权势案例的探测却变得越来越困难。另外,因为它们没有考虑 y 值与数据主体的模式的一致程度就直接根据 x 值来削减案例的权数,使得估计的效率退化(参见Krasker and Welsch,1982)。其他证据还暗示史威普估计量在误差分布不对称时不具备一致性(Carroll and Welsh,1988),这意味着它们对更常见的特异值出现在单侧的问题无能为力,而这正是本书主要关注的问题。

为了尝试解决这些问题,其他的GM估计程序则首先完全移除这些高度异常的案例,然后再对剩下的"好"案例应用M估计(Coakley and Hettmansperger,1993;Chave and Thomson,2003)。其中最值得一提的,应该算寇克里和赫特曼斯伯格的史威普一步估计(Schweppe one-step estimator,SIS),它由原始的史威普估计量扩展而来。这一估计量比原始估计量好的地方,在于它的杠杆权数考虑了观察案例与数据主体部分的一致性。换句话说,它考虑了观察案例是"好"的杠杆点还是"坏"的杠杆点,只给予后者较轻的权重。这样,在高斯—马

① 原书有误,权数的字母应该是 w,而非原书的 v。——译者注

科夫假定下，相对于 OLS，估计它有 95%的效率。

SIS 估计量初始的残差和残差尺度来自于崩溃点更高的回归，而不像此前发展出来的 GM 估计量一样取自 OLS 回归。它使用鲁索的 LTS 估计来获得初始估计，从而使得崩溃点 $BDP = 0.5$。这种方法也不同于马娄斯和史威普估计，它在纳入来自 LTS 回归的原初估计后，只经过一步而非迭代的方式就计算出最终的 M 估计（因此被命名为“一步估计”）。不过，尽管 SIS 估计量比其他 GS 估计量效率更高，在正态分布及大样本情况下甚至能够与 OLS 估计媲美，仿真研究显示，它们的效率在 n 很小时非常低（见 Wilcox，2005：438—440）。

S 估计量（S-Estimators）

为了应对 M 估计崩溃点过低的问题，Hampel（1975）建议考虑残差的尺度。按照这一思路，鲁索和约哈依（Rousseeuw and Yohai，1984，又见 Rousseeuw and Leroy，1987）提出了 S 估计量。S 估计是使得残差离散性最小化的解：

$$\min \hat{\sigma}(e_1(\hat{\beta}), \cdots, e_n(\hat{\beta})) \quad [4.18]$$

这显然可以和 OLS 进行比较，后者最小化的是残差的方差。因此，OLS 估计可以被看作是 S 估计的特例，一个不那么稳健的特例。而稳健 S 估计最小化的不是残差方差，而是残差尺度的某种稳健 M 估计：

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{\hat{\sigma}_e}\right) = b \quad [4.19]$$

其中 b 为常数，定义为 $b = E_\Phi[\rho(e)]$，而 Φ 代表标准正态分布。求等式 4.19 的导数，并解出下式的结果：

$$\frac{1}{n}\sum_{i=1}^{n}\Psi\left(\frac{e_i}{\hat{\sigma}_e}\right) = b \quad [4.20]$$

其中 Ψ 由某一合适的权数函数替代。和绝大多数 M 估计程序一样，Huber 权数函数或双权函数是经常被使用的。尽管 S 估计的崩溃点 $BDP = 0.5$，但代价是，它们的效率相对于 OLS 估计而言非常低（大概只有 30%）（Croux，Rousseeuw and Hossjer，1994）。

广义 S 估计量(Generalized S-Estimators)

克鲁克斯等(Croux et al. , 1994)提出了广义 S 估计(Generalized S-Estimates, GS 估计)以试图解决原版 S 估计量的低效率问题。这些估计量是通过寻找残差的尺度的某种 GM 估计量而产生出来的。GS 估计量的一个特例是最小四分位差估计量(least quartile difference estimator, LQD),这一估计量并行计算时用四分位差(interquartile range)来估计变量的尺度。LQD 估计量是这样定义的:

$$\min Q_n(e_1, \cdots, e_n) \qquad [4.21]$$

其中:

$$Q_n = \{|e_i - e_j|;\ i<j\}_{\binom{h_p}{2}:\binom{n}{2}} \qquad [4.22]$$

且:

$$h_p = \frac{n+p+1}{2} \qquad [4.23]$$

而 p 为模型中参数的个数。更简单地讲,这意味着 Q_n 是集合 $\left\{\left|e_i - e_j\right|;\ i<j\right\}$ 的 $\binom{n}{2}$ 个元素中的第 $\binom{h_p}{2}$ 个序次统计量(order statistic)。尽管这些估计量比 S 估计量更有效率,但带有"轻微增强的最差案例偏差(slightly increased worst-case bias)"(Croux et al. , 1994:1271)。

约哈依和扎马尔(Yohai and Zamar, 1988)的 τ 估计同样是通过将残差尺度的某种估计最小化来定义的,但其权重随着潜在的误差分布是自适应的,由此产生了崩溃点很高且效率很高的误差尺度估计。尽管如此,杠杆效应高的点并没有被考虑到,因此估计量的效率仍然受到阻碍。菲尔蒂(Ferretti et al. , 1999)曾试图利用广义 τ 估计(generalized τ estimates)来解决这一不足,它使用的权重考虑到了杠杆效应高的观察案例,处理方式和由 M 估计扩展而来的 GM 估计的很像。这种方法具有很高的崩溃点(高达 0.5)和相对于其他 GS 估计更

高的效率(尽管仍然只有 75%)。但是,相对于很多其他估计量,75%的效率仍然太低,这限制了 S 估计作为单独估计量的应用。不过,因为它们对特异值具有很强的抗扰性,S 估计在效率高得多的 MM 估计的运算中作用很大。

MM 估计量(MM-Estimators)

MM 估计最先由约哈依(Yohai, 1987)提出,现已变得越来越流行,或许可以说是目前使用最多的稳健回归技术。它们同时具有高崩溃点(50%)和良好的效率(在高斯—马科夫假定下,效率约为 OLS 估计的 95%)。它的名称中"MM"指的是这种估计使用了一个以上的 M 估计程序来计算最终的估计。与 M 估计的情况一样,迭代再加权最小二乘法(IRLS)被用来求解最终估计。程序如下:

(1) 系数 $\hat{\beta}^{(1)}$ 和对应残差 $e_i^{(1)}$ 的初始估计取自于具有高度耐抗性的回归(即崩溃点为 50%的回归)。只要这一估计量具有一致性,不一定需要有效率。因此,带 Huber 权数或双平方权数的 S 估计(可以被看作是 M 估计的一种形式)通常被用在这一阶段。[19]

(2) 第一步初始估计所得残差 $e_i^{(1)}$ 被用来计算残差尺度的某一 M 估计,$\hat{\sigma}_e$。

(3) 从第 1 步得到的残差 $e_i^{(1)}$ 的初始估计和从第 2 步得到的残差尺度 $\hat{\sigma}_e$ 的初始估计被用来进行加权最小二乘的第一次迭代,以确定回归系数的 M 估计:

$$\sum_{i=1}^{n} w_i(e_i^{(1)}/\hat{\sigma}_e)x_i = 0 \qquad [4.24]$$

其中 w_i 通常为 Huber 权数或双平方权数。

(4) 使用初始 WLS 估计(第 3 步)的残差计算出新的权重 $w_i^{(2)}$ 。

(5) 在保持第 2 步、第 3 步、第 4 步的残差的尺度不变的情况下,不断迭代直到收敛为止。

各种估计量的比较

表4.1总结了我们讨论过的多数估计量的一些稳健特性。给出的内容包括崩溃点，该估计量的影响函数是否有界，以及相对于OLS估计量的大概渐近效率。很明显，LAV和M估计量的崩溃点相对较低，受数据具体形式的影响大，这两种估计量的表现有时甚至不比OLS估计量好。哪怕一个特异案例，就能使这些估计变得毫无用处。有限影响R估计量同样好不了多少，其崩溃点小于 $BDP = 0.2$。与其他估计量相比，这几者几乎可以忽略不计，至少从单独使用(时的功效)来看是这样。

表4.1　各种回归估计量的稳健属性

估　计　量	崩溃点	有限影响	渐近效率
OLS	0	No	100
LAV	0	Yes	64
LMS	0.5	Yes	37
LTS	0.5	Yes	8
LTM	0.5	Yes	66
有限R估计	<0.2	Yes	90
M估计(Huber,双权)	0	No	95
GM估计(Mallows, Schweppe)	$1/(p+1)$	Yes	95
GM估计(SIS)	0.5	Yes	95
S估计	0.5	Yes	33
GS估计	0.5	Yes	67
广义估计	0.5	Yes	75
MM估计	0.5	Yes	95

我们还需要注意那些效率很低的估计量，例如LMS, LTS, LTM和S估计量。如果我们的目标在于保证耐抗性，而无需对总体进行推论，那这些估计量可能还是合适的。相反，如果对特异案例的性质不是很了解，那最好别使用这些估计量。盲目使用将会导致效率较低的估计，错过其他效率更高的可能

选择。例如，如果误差服从正态分布，那使用 OLS 估计量将会好很多。

M 估计尽管崩溃点很低，但效率高仍是一个很好的属性。将其与耐抗性更高的其他估计量组合使用时，新估计量将同时具备较高的抗异常值干扰能力和更高的效率。例如，如果 M 估计使用的初始残差来自于耐抗性高的 LTS 拟合，得到的就是 SIS GM 估计量，后者对残差特异案例及杠杆效应高的案例都具有很强的抗扰力，并能保持相对于 OLS 估计量 95%的效率。以 LMS 估计或 S 估计的残差来计算 M 估计，也能得到类似高效且稳健的 MM 估计量。

例子 4.1：仿真数据

现在我们回到最早在第 3 章介绍的那个包含各类特异值的仿真数据。我们对这 3 个“被污染”数据集都试探了 6 个不同的回归估计：一个 OLS 估计、一个 LAV 估计、一个 M 估计（使用的是 Huber 权数）、一个 GM 估计（确切地讲是 Coakley-Hettmansperger 估计量）、一个 S 估计，最后一个是 MM 估计。每个数据集的各种估计拟合线都呈现在图 4.1 中，图中标出了各“污染”案例。

首先来看垂直特异值的情况，不管使用何种回归方法，基本结论几乎都一样，至少从斜率系数来看是这样。事实上，除了 OLS 回归线在截距上偏离较远外，其他各方法的回归线几乎难分彼此。另外，尽管 OLS 截距稍微小于其他模型的截距——表明回归直线被拉向特异值——但也并非如此不同以至成为问题。存在“好”的杠杆点(B)时，各种估计彼此更为相像，除 LAV 估计线比其他直线稍高外，其他回归直线都重叠在一起。而在数据中存在“坏”的杠杆点(C)的情况下，各估计的差异较大，尽管最显著的差异在 OLS 估计和其他估计之间。如我们在第 3 章中看到的，OLS 回归直线被严重拉向特异案例。但是，没有一个稳健回归估计被特异值严重影响。因此，很清楚，在最后这种情形下应该选择更为稳健的方法。如果是前两种情况，该怎么办呢？（实际上）这两种情况下各种估计之间的差异很小。

为了回答这一问题，我们来看残差的分布，看 OLS 估计的精度是否受到了影响。回想一下就能知道，残差宽度较小时，OLS 估计的标准误是最小的。

图 4.2 呈现了残差的分布情况，表明垂直特异值给 OLS 估计的标准误带来了问题(图 A)。相反，“好”的杠杆点的出现并没有影响残差的良好表现(图 B)。将图 4.1 和图 4.2 的信息综合到一块，表明 OLS 估计只适合于带好杠杆点的数据。既然所有稳健估计量得出的结论大体一样，那么在其他两种情况下，像 MM 估计量这样有效率的估计量就是最好的选择。

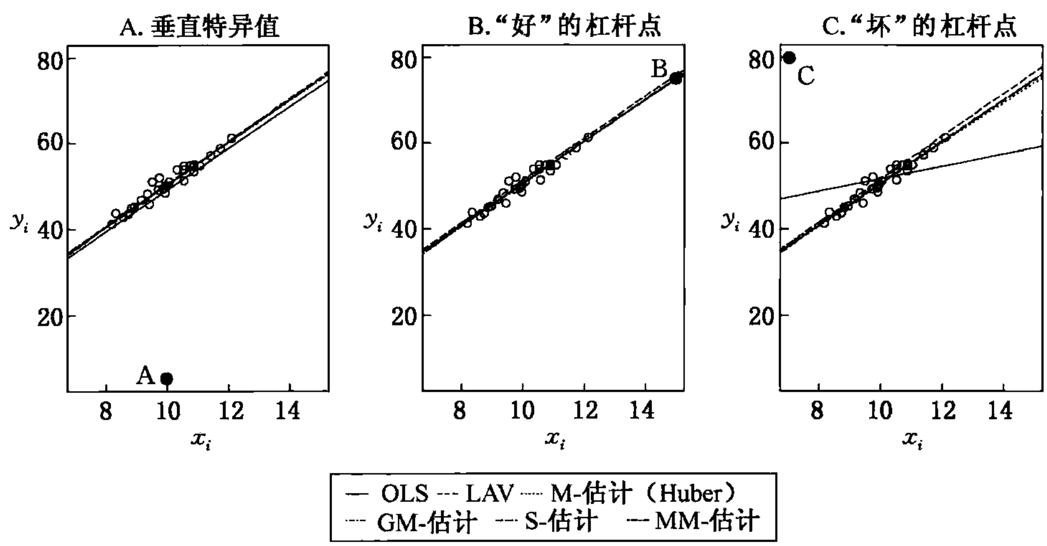


图 4.1　含 3 类特异值的人造数据的各种回归估计

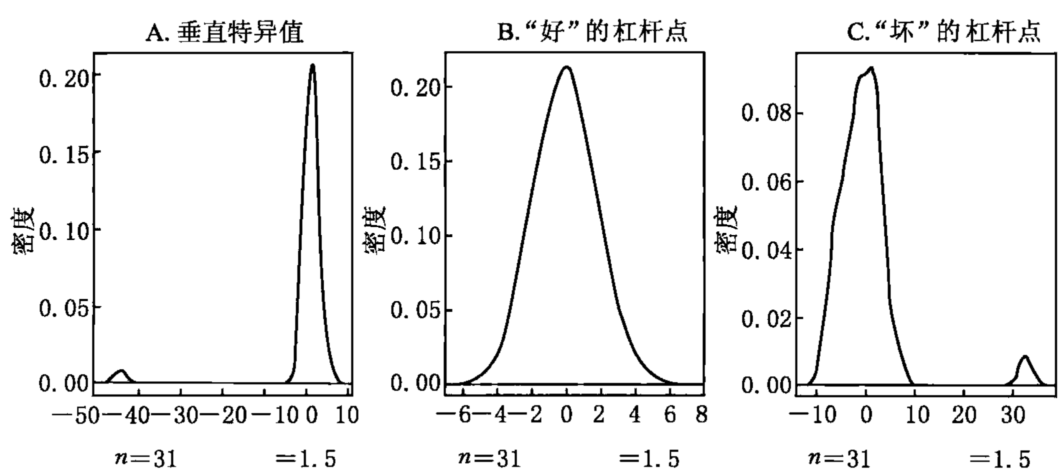


图 4.2　三个“被污染”数据集的 OLS 回归拟合所得残差的密度估计

例 4.2:预测公众态度的多元回归

下面我们回到跨国公众态度的数据,继续只关注那些新建立民主制度的国家和地区。在前面,我们使用了 OLS 模型,以人均 GDP 和基尼系数来预测公众态度。诊断和初步分析表明,模型在捷克及斯洛伐克被删除的情况下表现得更好(见表 3.2)。还记得如果包含所有案例,那么 OLS 模型中人均 GDP 的作用是显著的正作用 ($\hat{\beta}=0.0175$),而基尼系数的作用在统计上不显著 ($\hat{\beta}=0.00074$)。而在两个特异值被移除后,人均 GDP 的系数降到原来的 1/3,且统计上不再显著 ($\hat{\beta}=0.0063$);相反,基尼系数的斜率变成了原来的 7 倍并且统计显著 ($\hat{\beta}=0.00527$)。

表 4.2 给出了对上述相同数据拟合几个稳健回归的结果。尽管它们之间有一些小的差异,但 M 估计量、MM 估计量以及 GM 估计量的结果从人均 GDP 和基尼系数的影响来看非常相似。这些方法得到的结果也与删除两个特异案例之后的 OLS 回归的结果相似。LAV 回归也很好地发现了基尼系数和公众态度之间的关系,但得到的人均 GDP 的效果要小很多。不过,这一变量的作用即便是在移除了特异值的 OLS 回归中也统计不显著。总的来讲,各种稳健回归方法在应对权势案例上,做得比一般最小二乘回归好得多。

表 4.2 公众态度数据的各种稳健回归估计——新建立民主制度的国家和地区

	LAV 回归	M 估计 (Huber)	M 估计 (双权)	MM 估计	广义 M 估计 (Coakley-Hattmansperger)
截距	−0.079	−0.063	−0.091	−0.097	0.939
基尼系数	0.0045	0.0039	0.0049	0.0051	0.0041
人均 GDP(千美元)	0.0059	0.0089	0.0052	0.0057	0.0065
n	26	26	26	26	26

回归诊断再考察——稳健回归相关的特异值探测法

上面的讨论展示了稳健回归在限制特异案例影响上的优点。将它们作为

最终方法进行报告当然很合理，不过，它们也能被当作是初步分析的诊断工具（见 Atkinson and Riani，2000）。从这方面来讲，它们是对第 3 章讨论过的探测特异案例的传统技术的很好补充。

对于测量权势的指标（如 Cook's D）的批评之一，是它们不稳健。它们的计算基于样本均值及协方差矩阵，这意味着它们经常会遗漏特异值（见 Rousseeuw and van Zomeren，1990）。更具体地讲，Cook's D 容易受“掩盖效应（masking effects）”的影响，也就是说，如果存在一组权势点的话，它们会掩盖彼此的影响。我们已经知道，偏回归图在解决单个回归系数的掩盖效应上很有用处。而来自稳健回归的权数及残差信息则能在评估案例对回归的整体影响时帮助我们应对掩盖效应。

最终 IWLS 拟合权数标码图（Index Plots of Weight From the Final IWLS Fit）

将稳健回归当做诊断工具的直接方式之一，牵涉到最终 IWLS 拟合的权数。不过，不同模型的权数的意义不同，记住这一点非常重要。不同方法根据异常的类型（type of unusualness）给予观察案例的权数差别很大。对于 M 估计，关于权数我们唯一可以说的是，它们指示着在 OLS 拟合中残差的大小，也就是说案例是否为垂直特异案例。单独检查这些权数，并不能提供任何有关杠杆效应，因而也不能提供有关权势的信息，因为 M 估计在给出权重时根本不曾考虑这些因素。相反，GM 估计在削减权重时，同时考虑了案例在 OLS 回归中的残差和杠杆效应的大小，尽管此时对于权重的检查不允许我们对这两个方面进行区分。而来自 MM 估计的权数能够很好地指示案例对于回归整体的影响，因为这种估计的第一步具有很强的抗异常值扰动能力。

为评估来自各种稳健回归的权数的表现，将它们与对同一数据进行拟合的 OLS 回归的库氏距离（Cook's Distances）及其他特异值探测指标进行对比，能够提供很多有益的信息。表 4.3 就包括这些信息。在第 3 章中，通过库氏距离确定捷克和斯洛伐克为权势案例。所有稳健回归也都发现：对这两个特异案例，赋给它们的权数要相对小得多。换句话说，权数的大小表示着异常的程度：权

数越小,观察案例越异常。

表 4.3 来自 OLS 回归及稳健回归的诊断信息

国家或地区	OLS 诊断统计			稳健回归最终权数			
	Cook's D	帽子值	学生化残差	M 估计(Huber 权数)	M 估计(双平方权数)	MM 估计	GM 估计
亚美尼亚	0.0030	0.10	−0.27	1	0.81	0.87	0.35
阿塞拜疆	0.0011	0.10	−0.17	1	0.97	0.98	0.86
孟加拉国	0.0012	0.18	−1.13	1	1	1	1
白俄罗斯	0.0096	0.07	−1.63	1	1	1	1
巴西	0.0120	0.26	0.36	1	1	1	1
保加利亚	0.0011	0.07	0.21	1	0.82	0.88	0.38
智利	0.1350	0.24	1.14	0.72	0.55	0.72	0.18
中国	0.0005	0.073	0.13	1	1	1	1
克罗地亚	0.0073	0.063	−0.56	1	1	1	1
捷克共和国	0.3629	0.17	2.60	0.72	0	0	0.62
爱沙尼亚	0.0155	0.04	−1.00	1	0.85	0.89	0.40
格鲁吉亚	0.0011	0.07	−0.21	1	0.98	0.99	1
匈牙利	0.0273	0.098	−0.87	1	1	1	1
拉脱维亚	0.0093	0.07	−0.59	1	1	1	1
立陶宛	0.0040	0.04	−0.51	1	0.99	1	1
墨西哥	0.0003	0.17	0.06	1	1	1	1
摩尔多瓦	0.0038	0.112	0.295	1	0.96	0.97	0.84
尼日利亚	0.0504	0.14	0.95	1	0.89	0.93	0.47
罗马尼亚	0.0004	0.08	−0.11	1	0.91	0.94	0.55
俄罗斯	0.0189	0.09	−0.76	0.87	0.68	0.77	0.27
斯洛伐克	0.6990	0.17	4.32	0.17	0	0	0.05
斯洛文尼亚	0.2691	0.24	−1.66	1	0.99	0.99	1
中国台湾地区	0.1296	0.15	−1.50	1	0.96	0.97	0.80
土耳其	0.0011	0.05	0.26	1	0.89	0.93	0.46
乌克兰	0.0024	0.10	−0.26	1	0.83	0.88	0.38
乌拉圭	0.0059	0.07	0.46	0.95	0.65	0.78	0.24

来自表 4.3 的证据激发了我们对稳健回归权数制作带标签散点图的想法,就像对 Cook's D 所做的那样。图 4.3 中即做了这样的事。虽然说三种稳健回归方法都发现了这两个问题最严重的观察案例,但 GM 估计还给了另外 9 个案例小于 0.5 的权重,而其余方法给予其他任何案例的权数都不小于 0.7。前面已经说过,GM 估计的独特性,是因为它同时考虑了残差和杠杆效应的规模。

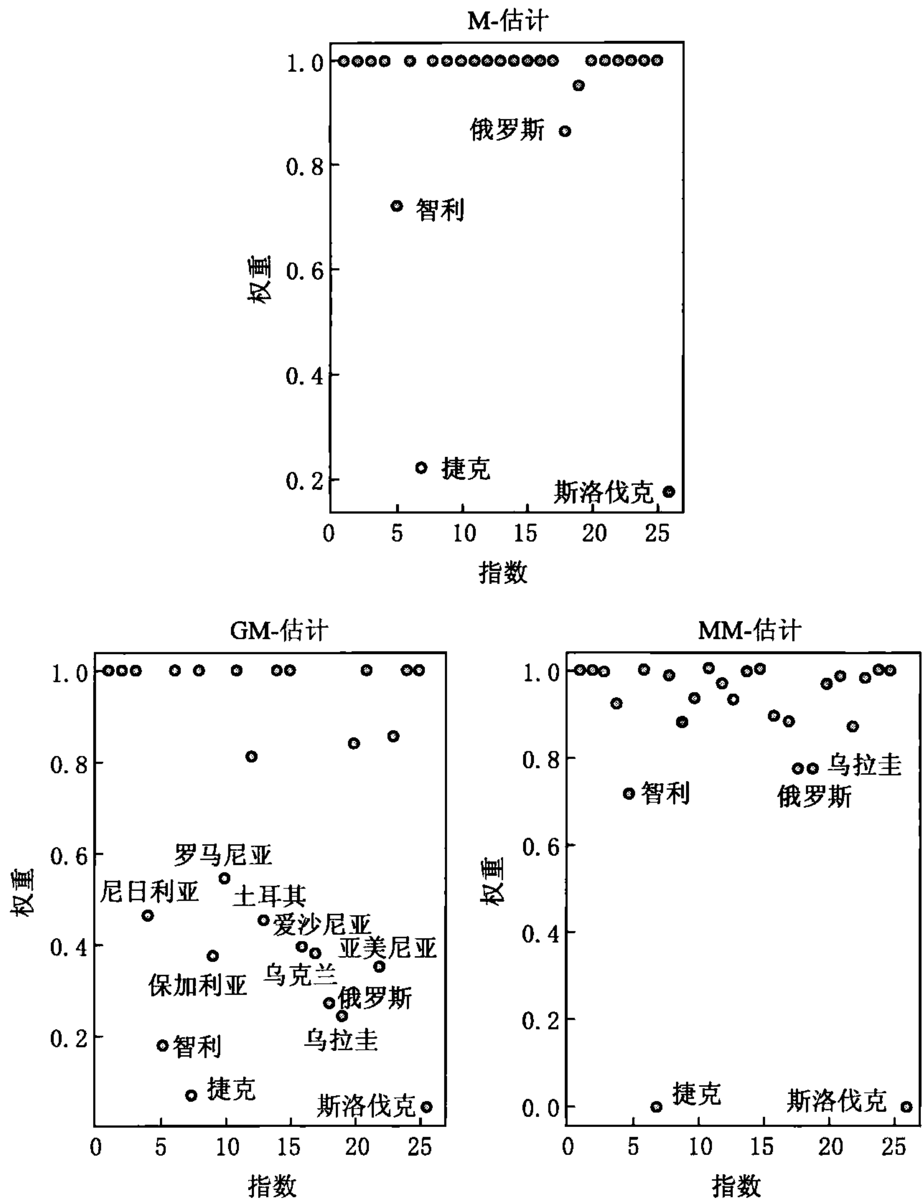


图 4.3　各种稳健回归估计 IWLS 拟合的最终权数标码图

RR 图（“残差—残差”图）RR-Plots（“Residual-Residual” Plots）

根据鲁索和佐默伦·冯（Rousseeuw and van Zomeren，1990:637）的观点，在诊断异常值方面，稳健回归残差比 OLS 残差要好得多，因为 OLS 回归“试图产生正态模样的残差，即使数据本身并不正态”。根据这一思想，图基（Tukey

1991)绘制出了 RR 图(“残差—残差”图),这一名字称呼的是一种散点图矩阵,它包括用 OLS 估计的残差分别对其他几种不同稳健回归的残差做出的多个散点图。如果 OLS 假定完全满足,那么 OLS 残差和其他所有稳健回归的残差之间将存在完美的正相关,且斜率等于 1[被叫做“同一线(identity line)”]。令来自第 j 个回归估计 $\hat{\beta}_j$ 的第 i 个残差为 $e_{ij}=y_i-x_i^T\hat{\beta}_j$,那么

$$\|e_i(\hat{\beta}_1)-e_i(\hat{\beta}_2)\|=\|\hat{y}_i(\hat{\beta}_1)-\hat{y}_i(\hat{\beta}_2)\|=\|x_i^T(\hat{\beta}_1-\hat{\beta}_2)\|$$
$$\leqslant\|x_i\|(\|\hat{\beta}_1-\beta\|+\|\hat{\beta}_2-\beta\|) \qquad [4.25]$$

这意味着:如果回归假定成立,随着 n 趋近于∞,围绕着同一线的散点将越来越紧。但如果存在特异值的话,斜率将不等于 1,因为 OLS 回归无法抵抗它们,而稳健回归却可以。

公众态度数据的 RR 图呈现在图 4.4 中。虚线为同一线;实线为纵坐标所

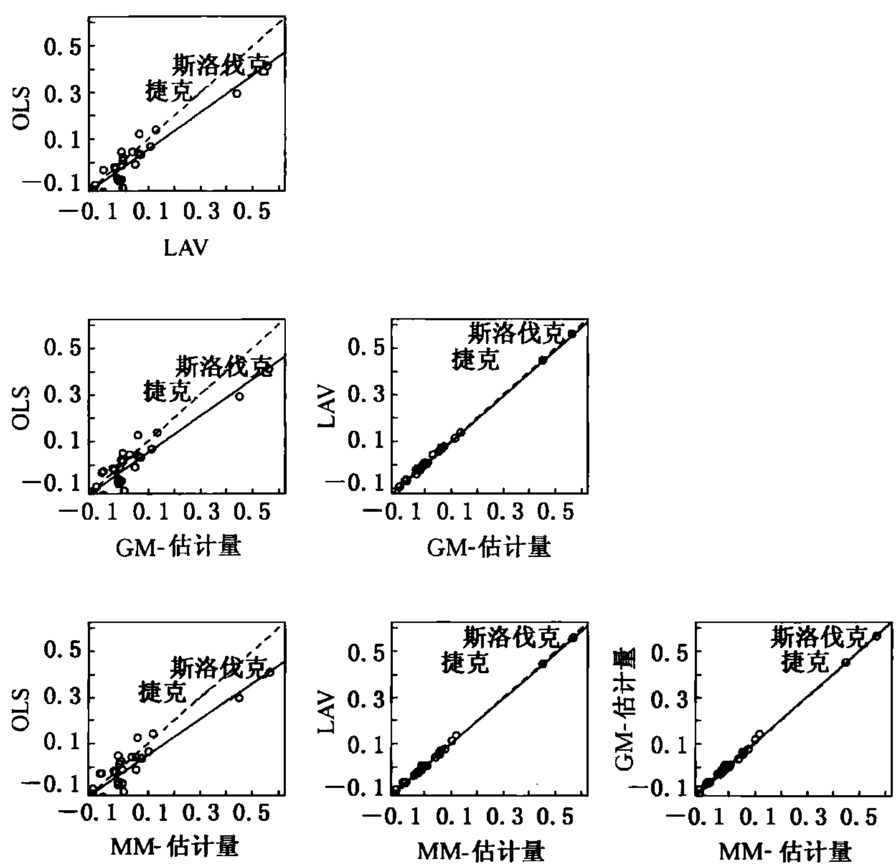


图 4.4 RR 图——以人均 GDP 及基尼系数对公众态度进行回归(新建民主国家及地区)

示回归方法的残差对横坐标所示回归方法的残差的回归直线。第一列的几个图是最重要的，因为它们呈现了 OLS 残差对其他各种稳健回归残差的回归情况。所有这些图中的两条直线相差得都很远，这一事实表明，OLS 估计受到了特异值的严重影响。捷克和斯洛伐克在 OLS 回归中的残差小得多，意味着它们的权势非常大。再看其他的图，可以发现各种稳健回归的残差彼此十分相像，尤其是 MM 估计和 GM 估计，两者的残差几乎完全相等。

稳健距离(Robust Distances)

我们也可以考虑只与一种稳健回归有关的诊断方法。例如鲁索和冯・佐默伦(Rousseeuw and van Zomeren, 1990)认为将稳健残差对稳健距离作图以探测复奇异点(multiple outliers)比传统方法更好，其中稳健距离建立在马氏距离(Mahalanobis distance)基础之上，但定义时使用的是稳健协方差矩阵(关于这一话题的争论还可以参见 Cook and Hawkins, 1990; Ruppert and Simpson, 1990；及 Kempthorne and Mendel, 1990)。因为这些诊断并不关系效率问题，所以耐抗性极高的 LMS 或 LTS 回归的残差被使用得最多。

所谓马氏距离测量的是一个观察案例 x_i 相对于由数据集 X 定义的所有数据点的中心的距离。定义如下：

$$\mathrm{MD}_i = \sqrt{(x_i - \bar{x})\mathrm{cov}\,(X)^{-1}\,(x_i - \bar{x})^T} \qquad [4.26]$$

其中 $\bar{x}$ 为 X 的重心，cov(X)为样本协方差矩阵。由于特异值能够影响均值和协方差矩阵，因此它们不一定能够通过 MD_i 被发现。因此鲁索和冯・佐默伦的稳健距离 RD_i(1990)在定义时将 cov(X)和 $\bar{x}$ 替换为更为稳健的来自于最小体积椭圆体估计量(minimum volume ellipsoid estimator)的中心和协方差矩阵(更详细的内容请看 Rousseeuw, 1985)。通常认为 $|e'| \geqslant 2.5$ 的标准化稳健残差是有问题的①。类似的，如果 RD_i 大于自由度与模型估计的参数个数相等的卡方分布的 0.975 分位点取值(percent point)，那么该稳健距离被认为有过高的

---

① 本句中的公式原书可能有错，绝对号应该加在变量上。——译者注

杠杆效应。

公众态度数据的鲁索和冯·佐默伦的回归诊断图见图 4.5。图中和稳健距离相对的是来自 LTS 回归的标准化残差。尽管从稳健距离来看,没有一个案例有异常高的杠杆效应,但稳健残差显示有 3 个案例是特异值。与截至目前为止我们所做的其他分析一致,斯洛伐克和捷克是其中之二。第 3 个案例是智利,刚好超过了拇指规则定义的分界线一点点。

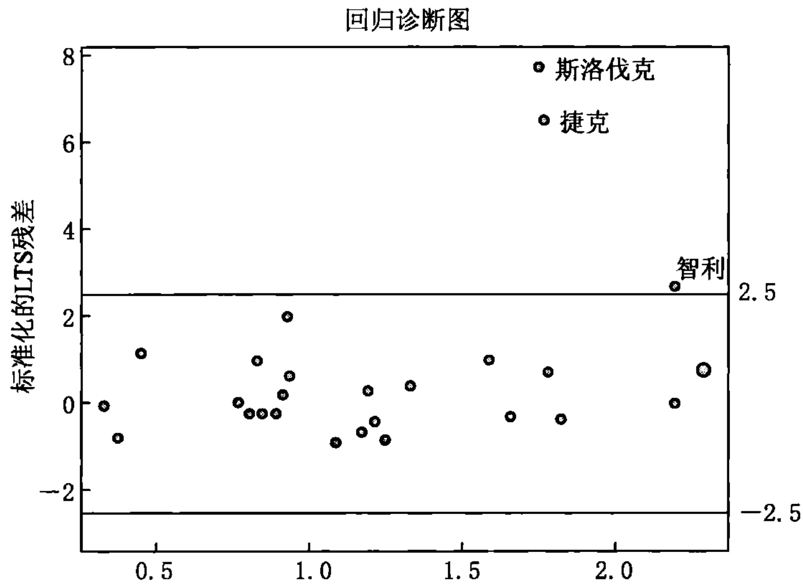


图 4.5　稳健残差(来自 LTS 估计)对稳健距离的散点图

和上面讨论过的方法一样,用来甄别特异案例的传统诊断图(第 3 章讨论过)也能扩展到稳健回归模型上。由于它们通常的解释方式和在 OLS 回归中的一样,这里就不再讨论了。关于这些诊断的更多信息,请看麦基恩和希哲的作品(McKean and Sheather, 2000)。与稳健回归相关的其他技术还可以看冯荣锦或佩纳及约哈依的著作(Fung, 1990; Pena and Yohai, 1999)。

第 5 章 | 稳健回归的标准误

分析标准误(analytical standard error)很容易就能从某些类型的稳健回归中计算出来,但并非都是如此。[20]不过,即使能够计算分析标准误,对小样本来说它们也很不可靠。因此,常常需要使用自助法(bootstrapping)来计算标准误。所以这一章开始将简要讨论渐近标准误(asymptotic standard errors),然后再探索各种类型的自助标准误及置信区间。

稳健回归估计量的渐近标准误(asymptotic standard errors)

分析标准误可以在 S 族和 M 族估计量(包括广义形式及 MM 估计量)中得到。这些估计量的渐近标准误(ASE)都是由与最终 IRLS 拟合的系数对应的渐近协方差估计矩阵的对角线元素的平方根给定的,$V(\hat{\beta}) = s_e^2 \, (X^T W X)^{-1}$ (Draper and Smith, 1998:575;还可参看 Hill and Holland, 1997 以及 Birch and Agard, 1993)[21],其中 W 为最终权数矩阵,s_e^2 为残差方差。残差的方差是这样定义的:

$$s_e^2 = \frac{1}{n-p} y \left[W^{-1} - W^{-1} X \, (X^T W^{-1} X)^{-1} \, X^T W^{-1} \right] y \qquad [5.1]$$

其中 p 为模型中的参数个数。这些估计结果在大多数统计软件的回归输出中都能找到。

如果样本规模 n 相对于要估计的参数个数足够大,渐近标准误(ASE)被认为是可靠的(Yohai, 1987)。但如果 n 很小(比方说小于 40),则难以信任(Li, 1985; Huber, 2004:164)。其他证据也表明:随着杠杆案例的增加,渐近标准误

的可靠性将会衰减(Stromberg, 1993)。因此通常建议只在样本较大时才使用渐近标准误(ASE)。而样本规模很小时,则可通过自助法获得另一种标准误或置信区间。[22]

自助标准误(Bootstrapped Standard errors)

自助法首先由埃夫戎(1979;又见 Efron, 1981)提出,它可以用来计算那些渐近标准误不那么容易求导或者作为使用条件的必要假定被违背的统计量的标准误及置信区间。自助标准误是通过反复从原始样本中抽样算出来的。尽管在 OLS 假定满足时自助回归并不可取——在这种情况下,OLS 估计是效率最高的无偏估计[23]——但如果假定没有得到满足,那么自助法将变得很有用。对稳健回归而言,当样本规模很小时,自助法最有用,因为在这种情况下 ASE 不足为信。稳健回归的自助回归可以通过两种方式实现:随机 x 自助法(random-x bootstrapping)或固定 x 自助法(fixed-x bootstrapping)。

随机 x 自助法

随机 x 自助法是对数据集的观察案例进行再抽样。也就是说,它是从数据矩阵中抽取数据行。当回归量随机时(regressors are random)——即每一个新的随机样本产生的结果应该不同——这是一种合适的做法,在大规模抽样调查数据中就是这样(Mooney and Duval, 1993:17)。假定一个只有一个自变量 x_i 和一个因变量 y_i 的数据集。获取随机 x 自助标准误的步骤非常简单:

(1) 从原始样本数据(x, y)中回置性地随机抽取出 $B = 1, \cdots, R$ 个规模为 m 的样本(x^*, y^*)。这 R 个样本被叫做自助样本,B。对于小样本,通常需要抽取 $R = n^n$ 个自助样本。对大样本来讲,抽取 n^n 个样本是不实际的(如,运算 1000^{1000} 个自助样本将会是一件愚蠢的事情),但 1000 个自助样本通常被认为是可以接受的。

(2) 对 R 个自助样本中的每一个都计算稳健回归估计 $\hat{\beta}_0^*$ 和 $\hat{\beta}_1^*$。

使用 $\hat{\beta}_0^*$ 和 $\hat{\beta}_1^*$ 的经验分布计算 $\hat{\beta}_0^*$ 和 $\hat{\beta}_1^*$ 的标准误，使用的推论方法和用于来自大总体的随机样本的经典方法一样。换句话说，自助标准误是由自助样本统计量的分布计算而来，而非来自该统计量未知的抽样分布。

固定 x 自助法

固定 x 自助法是当解释变量（即模型矩阵 X）被假定固定不变时，随机 x 自助法的一种可供选择的适当替代物。这种方法比随机 x 自助法相对复杂一些，它是对回归模型的残差进行再抽样（resamples the residuals）而不是对观察案例本身再抽样。它的程序如下：

（1）将初始稳健回归的预测值 $\hat{y}_i$ 作为自助回归反应变量的期望值。

（2）和通常一样，从回归模型计算出残差：$e_i = y_i - \hat{y}_i$。

（3）从残差 e_i 中回置性地随机抽取出规模为 n 的样本 $B=1, \cdots, R$。和在随机 x 自助法中一样，自助 1000 次是通常的选择。再抽样得到的残差叫作 $\hat{e}_{B_i}^*$。

（4）将再抽样得到的残差 $\hat{e}_{B_i}^*$ 加到回归预测值上，得到固定 x 自助样本，$y_{\hat{\beta}_i}^* = \hat{y}_i + \hat{e}_{B_i}^*$，形成 R 套自助预测值。

（5）将 R 套自助预测值中的每一套都对固定模型矩阵 X 做回归，得到 R 套回归系数。

（6）从自助法复制得到的回归系数的经验分布产生出置信区间（或标准误）。

置信区间的构造(Constructing Confidence Intervals)

不管是使用随机 x 自助法还是固定 x 自助法，自助回归系数的平均值都是：

$$\bar{\hat{\beta}}^* = \hat{E}^*(\hat{\beta}^*) = \frac{\sum_{B=1}^{R} \hat{\beta}_B^*}{R} \qquad [5.2]$$

$\hat{\beta}^*$ 估计的自助方差为：

$$\hat{V}^*(\hat{\beta}^*) = \frac{\sum_{B=1}^{R}(\hat{\beta}_B^* - \bar{\hat{\beta}}_B^*)^2}{R-1} \qquad [5.3]$$

有三种类型的置信区间通常被考虑:(1)正态论置信区间,(2)分位置信区间,(3)误差修正分位置信区间。通过作图对自助分布进行检查有助于决定使用何种置信区间。

如果自助抽样分布的均值无偏,也就是$\bar{\hat{\beta}}^* = \hat{\beta}$,那可以考虑正态论区间或分位区间。顺着很多统计量都服从渐近正态分布这一想法,当自助抽样分布为渐近正态分布时,正态论区间(normal-theory intervals)也就具有合理性。一个 $100(1-\alpha)\%$ 的置信区间按照标准方式计算,如下:

$$\beta = \hat{\beta} \pm z_{\alpha/2}\hat{SE}^*(\hat{\beta}^*) \qquad [5.4]$$

其中标准误的自助估计,$\hat{SE}^*(\hat{\beta}^*)$,是自助抽样分布的标准差。

如果自助统计量的均值无偏而自助抽样分布非正态的话,选用分位区间更合适。这种区间的计算,首先需要将自助统计量从小到大排列:$\hat{\beta}_{(1)}^* \leqslant \hat{\beta}_{(2)}^* \leqslant \cdots \leqslant \hat{\beta}_{(R)}^*$。然后将置信区间限定在自助抽样分布的 $\alpha/2$ 和 $1-\alpha/2$ 分位,$\hat{\beta}_{(\alpha/2)}^* < \beta < \hat{\beta}_{(1-\alpha/2)}^*$,其中 $100(1-\alpha)\%$ 为设定的置信水平。

正态论置信区间及分位置信区间在自助估计无偏时效果良好,大样本时通常都是这种情况。如果估计量偏差相当大——小样本情况下很常见——那就必须进行调整。为达到这个目的,偏差修正置信区间(bias corrected [BC] confidence intervals)使用了一个包含两个纠正因子,Z 和 A 的正态化转换。Z 是这样定义的:

$$Z = \Phi^{-1}\left[\frac{\#_{b=1}^{R}(T_b^* \leqslant T)}{R+1}\right] \qquad [5.5]$$

其中 Φ 为标准正态密度函数,$\#_{b=1}^{R}(T_b^* \leqslant T)/(R+1)$ 是初始样本估计 $\hat{\beta}$ 下自助复制系数的修正比例(the adjusted proportion of the bootstrap replicates below the original sample estimate)。修正因子 A 定义如下:

$$A = \frac{\sum_{i=1}^{n} (T_{(-i)} - \overline{T})^3}{6\left[\sum_{i=1}^{n} (T_{(-i)} - \overline{T})^2\right]^{3/2}} \qquad [5.6]$$

其中 $T_{(-i)}$ 代表移除第 i 个案例的折刀法再抽样[24]的 T 值，是 n 个折刀值(jacknife values)的平均数，也就是，$\overline{T} = \sum_{i=1}^{n} T_{(-i)}/n$。偏差修正置信区间(BC confidence interval)的上限和下限再按照下面的方式计算出来：

$$BC_{lower} = \Phi\left[Z + \frac{Z_{-z_{\alpha/2}}}{1 - A(Z_{-z_{\alpha/2}})}\right]$$

$$BC_{upper} = \Phi\left[Z + \frac{Z_{+z_{\alpha/2}}}{1 - A(Z_{+z_{\alpha/2}})}\right] \qquad [5.7]$$

关于 BC 置信区间的更多细节，可以看埃夫戎和提布希拉尼的合著(Efron and Tibshirani，1993：第 14 章)以及戴维森和辛克利的书(Davidson and Hinkley，1997：103—107)。

例 5.1：民主对收入不平等及人均 GDP 对公众态度的影响的影响

继续使用跨国公众态度的数据，这次我们对以基尼系数、人均 GDP 和民主制度解释公众态度的模型进行探索。初步的分析已经显示基尼系数和民主之间存在交互作用，因此我们在模型中纳入了这一交互项。诊断分析虽然在这里没有呈现出来，但我们发现：捷克和斯洛伐克仍然对回归估计有着异乎寻常的影响，这意味着稳健回归是 OLS 回归的很好的替代者。各种模型的回归系数及标准误都在表 5.1 中报告出来了。表中包含的估计分别来自于包含所有案例的 OLS 回归，两个特异值被移除的 OLS 回归，M 估计，以及 MM 估计。

尽管报告了稳健回归的渐近标准误，但它们并不足信，因为样本的规模太小 ($n = 48$)，因此，这里也报告了自助标准误。观察案例代表的是国家或地区——因而模型矩阵可以合理地被认为是固定的——所以这里使用固定 x 再抽样来计算自助置信区间。我们从初始稳健回归的残差中抽取了 1000 个自助样本。图 5.1 只给出了民主系数 M 估计的自助(残差)诊断图——一个直方图

和一个自助分布的分位比较图(quantile comparison plot of the bootstrap distribution)。直方图中间的垂直虚线代表自助估计的均值,可以看出它和稳健回归的系数几乎完全相等($\hat{\beta}=0.374$)。也就是说,自助估计的结果是无偏的。而分位比较图表明分布稍偏重尾,但大体还是正态分布,这告诉我们应该使用分位置信区间。其他系数的回归诊断结果类似,因此,表 5.1 中所有自助标准误都是基于分位置信区间的结果。

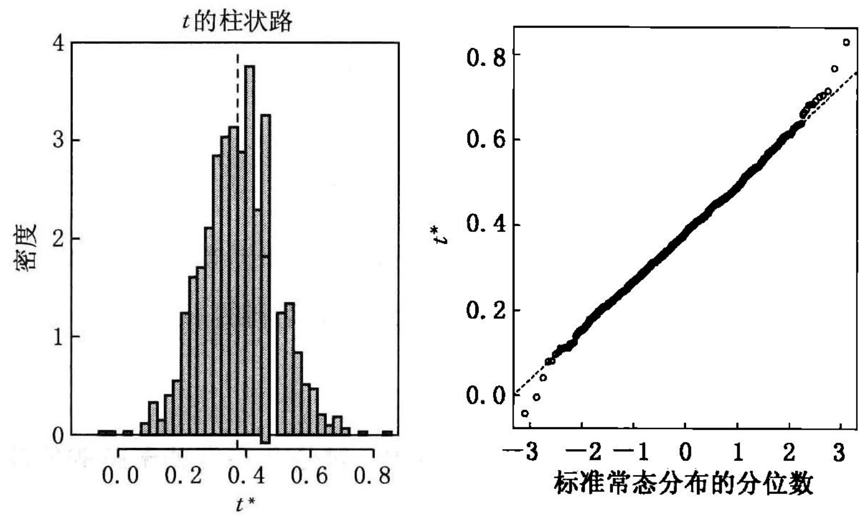


图 5.1 来自 M 估计稳健回归的民主系数的自助复本诊断图

表 5.1 中最显眼的是 OLS 估计中两个特异值的持续有害影响。这两个特异案例存在时,OLS 回归系数没有一个统计显著。当它们被删除时,除了人均 GDP 仍然不显著外($p=0.82$),其他系数都显著提高,并且都变得统计显著了(最大的 $p=0.0032$)。两个稳健回归方法的差异也很明显。与前面的例子相反①,在这里,我们发现:M 估计比 MM 估计更容易受特异观察案例影响。因为 M 估计在降低观察案例权重时仅考虑它们的残差,而未考虑杠杆效应。但捷克和斯洛伐克除了有大残差之外,还有很大的帽子值。这种过高的杠杆效应给 M

① 原书为"contrary to the previous example"。实际上,M 估计未考虑杠杆效应,不具有界影响函数,而 MM 估计具有有界的影响函数,比 M 估计更稳健,尽管前面的例子中 M 估计和 MM 估计相差不大。——译者注

表 5.1 对公众态度进行预测的 OLS 回归及稳健回归的估计及其标准误(全部国家和地区)①

	OLS(全部案例)		OLS(特异值已移除)		M 估计(Huber 权数)			MM 估计		
	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\beta}$	ASE($\hat{\beta}$)	自助标准误(固定 X) $R=1\,000$	$\hat{\beta}$	ASE($\hat{\beta}$)	自助标准误(固定 X) $R=1\,000$
截距	1.162	0.098	0.939	0.063	1.019	0.064	0.061	0.952	0.061	0.061
民主	0.114	0.182	0.467	0.115	0.374	0.119	0.112	0.475	0.113	0.119
基尼系数	−0.000 6	0.002	0.004 9	0.001 6	0.002 9	0.001 6	0.001 5	0.004 6	0.001 5	0.001 5
人均 GDP/1 000	0.003 9	0.004	0.000 5	0.002	0.001 0	0.002 5	0.002 3	0.000 2	0.002 3	0.002 4
民主×基尼系数	−0.002 8	0.005	−0.010	0.002	−0.008	0.000 3	0.003 0	−0.011	0.002 9	0.003 0
n	48		46		48			48		

① 原书表格中最后一行案例数 n 排布错误,这里已经改正。——译者注

估计造成了麻烦，尤其是对基尼系数，M 估计中这一变量的效应要比不带特异案例的 OLS 估计和 MM 估计的小得多。而有限影响 MM 估计与不带两个特异案例的 OLS 估计结果几乎相同，这表明前者给予这两个特异案例的权数为 0。第三点值得注意的，是渐近标准误和自助标准误很相似。没有一个系数因为标准误的差异而严重影响到显著性的检验，从而表明渐近标准误用在这个例子里还不错。

为了更清楚地查看这些模型估计结果的差异，图 5.2 给出了 OLS 回归和 MM 回归的拟合值，它们呈现了民主和基尼系数之间的交互作用。要想计算这些预测值，需要将回归方程中的人均 GDP 设定为变量均值，但允许民主和基尼系数在各自的值域范围内变动。换句话说，这些回归拟合线呈现了一个人均 GDP 为典型值的国家的基尼系数和民主制度的联合效应。考虑到 OLS 回归中没有一个系数统计显著，因此拟合图呈现的效应模式并不明显也就不奇怪了。相反，在 MM 估计中，交互作用明显很强。对于传统的民主国家，随着收入不平等(基尼系数)的提高，公众态度倾向于报酬平等的程度有所下降。对于建立民主制度不久的国家和地区，情况则恰恰相反。尽管方向不同，但基尼系数对新的和传统的民主体都有很强的影响。

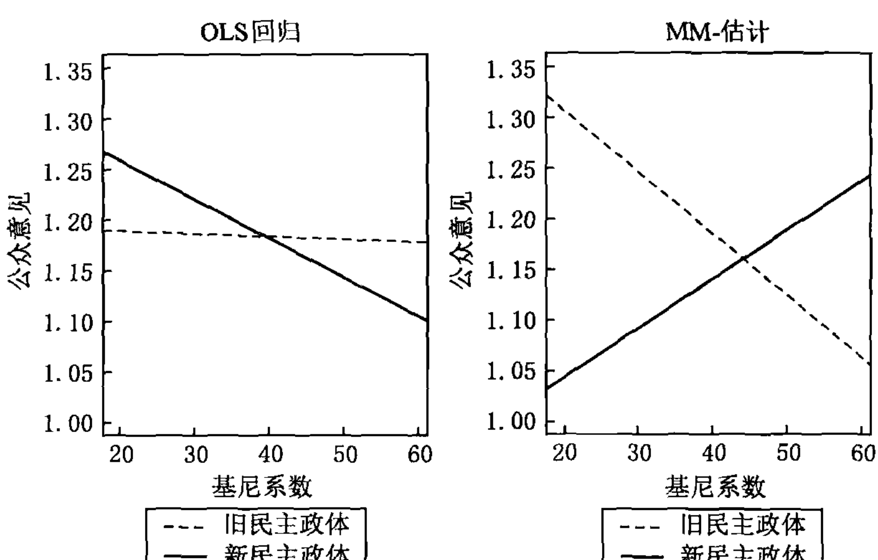


图 5.2 民主与基尼系数的交互作用对公众态度的影响的展示——OLS 回归估计(含特异值)与 MM 估计

第 6 章 | 广义线性模型中的权势案例

广义线性模型(Generalized Linear Models,简称 GLM)由一般线性模型扩展而来,以适应分布非正态的因变量,包括那些非连续因变量。本章开始部分将概要描述 GLM 模型。然后简单讨论在 GLM 中探测特异案例的诊断方法。最后将对稳健广义线性模型进行介绍,并为 logistic 模型和泊松模型提供了实际的例子。

广义线性模型

这里我仅提供 GLM 的基本描述,并强调那些对于理解稳健广义线性模型所必需的信息。关于 GLM 的更丰富和具体的描述,可以看麦卡拉和内尔德(McCullagh and Nelder, 1989)关于这一话题的经典著作(关于 GLM 的一般论述,还可以看其他著作如 Dobson, 1990; Fahrmeir and Tutz, 2001; Lindsey, 1997)。专门面向社会科学家讨论 GLM 的作品,在本丛书中就有 3 本(Gill, 2001; Dunteman and Ho, 2005; Liao, 1994)。

还记得线性模型是这样表示的:

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \qquad [6.1]$$

其中 y 被假定与 x 线性相关,而误差项被假定相互独立,方差固定,并且服从正态分布。换句话说,线性模型代表的是给定各 x 的情况下 y 的条件均值:

$$\mu_i = \sum_{j=1}^{k} x_{ij}\beta_j \qquad [6.2]$$

广义线性模型放宽了这些假定，以对服从指数分布的因变量的条件均值进行预测，它的一般形式如下：

$$f(yi;\theta i;\varphi)=\exp\left[\frac{y\theta-b(\theta)}{a(\varphi)}+c(y,\varphi)\right] \quad [6.3]$$

其中 θ 为代表位置估计量的标准参数(canonical parameter)，而 φ 为代表尺度的离散参数(dispersion parameter)。也就是说，GLM 允许 y 的分布服从众多不同的指数族外形：

$$yi\left|xi\sim\begin{cases}\text{高斯分布}\\\text{二项分布}\\\text{泊松分布}\\\text{伽马分布}\\\text{等等}\end{cases}\right.$$

具体的指数族由方程 6.3 中的 a, b, c 确定。

线性关系假定在 GLM 中仍然存在，但是是相对于线性预测值(linear predictor) η 而非 y 本身：

$$\eta_i=\sum_{j=1}^{k}x_{ij}\beta_j \quad [6.4]$$

也就是说，等式 6.3 中的标准参数 θ 取决于线性预测值。更准确地讲，因变量的条件均值 μ_i 是通过某种转换与线性预测值连接在一起的，这种转换被称为连接函数(link function) $g(\cdot)$：

$$g(\mu_i)=\eta_i \quad [6.5]$$

连接函数必须是单调函数(monotonic)且可导(differentiable)，能够取任意值(正的或负的)，以保证 η 线性地取决于解释变量。当连接函数被设定为恒等连接(identity link)而分布被设定为高斯族(Gaussian family)时，拟合的模型即为 OLS 回归。其他任何连接函数得到的都是因变量 y_i 的期望值与自变量 x_{ij} 之间的非线性关系。表 6.1 呈现了 GLM 框架涵盖的一些重要分布族及对应的连接函数。

表 6.1 重要的指数分布族及相应的连接函数

分　布	μ 的值域	连接函数, $g(\mu)$	
正态分布	$(-\infty, +\infty)$	恒等函数	$g(\mu)=\mu$
二项分布	$(0, 1)$	Logit 函数	$g(\mu)=\log[\mu/(1-\mu)]$
	$(0, 1)$	Probit 函数	$g(\mu)=\Phi^{-1}(\mu)$
泊松分布	$(0, \infty)$	对数函数	$g(\mu)=\log(\mu)$
伽马分布	$(0, \infty)$	倒数函数	$g(\mu)=\mu^{-1}$
	$(0, \infty)$	对数函数	$g(\mu)=\log(\mu)$

GLM 广义线性模型的最大似然估计，是通过把方程 6.3 看作是参数 β 的某个函数求解出来的。通常来说，这意味着把和 β 有关的对数似然函数(log-likelihood function)最大化：

$$
\begin{aligned}
l(\beta) &= \log L(\beta) = \log \prod_{i=1}^{n} f(y_i;\ \mu_i) \\
&= \log \prod_{i=1}^{n} f(y_i;\ xi,\ \beta) = \sum_{i=1}^{n} \log f(y_i;\ x_i,\ \beta)
\end{aligned} \qquad [6.6]
$$

最大似然估计结果可以用 Newton-Raphson 法或迭代再加权最小二乘法得到(见 Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989)。在对 GLM 模型的 IRLS 估计中，因变量不是 y 本身，而是调整因变量 z(adjusted dependent variable)，它是适用于 y 的连接函数的线性化形式。我们先来定义第一次迭代的线性预测值：

$$
\underset{(n\times 1)}{\hat{\eta}^{(0)}} = \underset{(n\times p)}{X^T}\ \underset{(p\times 1)}{\beta^{(0)}} \qquad [6.7]
$$

并同时从 $g^{-1}(\hat{\eta}^{(0)})$ 中得到初始拟合值 $\hat{\mu}^{(0)}$。然后我们定义 z 如下：

$$
z^{(0)} = \hat{\eta}^{(0)} + \left(\frac{\partial \eta}{\partial \mu}\bigg|_{\hat{\mu}^{(0)}}\right)(y-\hat{\mu}^{(0)}) \qquad [6.8]
$$

IRLS 中使用的二次加权矩阵(quadratic weight matrix)由下面的表达式确定：

$$
\mathrm{W}_{(0)}^{-1} = \left(\frac{\partial \eta}{\partial \mu}\bigg|_{\hat{\mu}^{(0)}}\right)^2 \mathrm{V}(\mu)\bigg|_{\hat{\mu}^{(0)}} \qquad [6.9]
$$

其中$V(\mu)$是在$\hat{\mu}^{(0)}$时给定的方差函数。而z和$W_{(0)}$都取决于当下的拟合值，因此需要一个迭代过程，以求出最终解。我们首先将$z^{(0)}$对x做回归，相应权数为$W_{(0)}$，得到新估计的回归系数$\hat{\beta}^{(0)}$，以及由此而产生的线性预测值的新估计。通过使用新的z和W估计值，估计程序将持续下去，直到收敛为止，得到如下一般形式的标准方程：

$$\hat{\beta} = (X^T W X)^{-1} X^T W z \qquad [6.10]$$

其中z代表经连接函数变换的调整因变量，W为最终权数矩阵。GLM模型还能被拟似然估计(quasi-likelihood estimation)进一步扩展，这种估计在通常的连接函数设定基础上，允许对离散参数φ而非y的整个分布进行设定(更多细节请看Wedderburn，1974)。

模型的偏差(Deviance)与最小二乘回归中的残差平方和的角色类似，它将当下研究的模型与相同数据的饱和模型β_S进行比较。对于案例数为n的数据，带n个回归系数的饱和模型能够完美地拟合该数据，意味着它达到了最高可能达到的似然值。这一饱和模型的似然值为比较其他非饱和模型的似然值提供了基准。偏差测量的正是所设模型与饱和模型的差异程度。更确切地讲，它等于饱和模型的对数似然值与当下研究的模型的对数似然值之差的两倍：

$$\begin{aligned} D(\beta, y) &= 2[\log L(\beta_S)] - 2[\log L(\beta)] \\ &= -2[\log L(\beta)] \end{aligned} \qquad [6.11]$$

偏差在模型拟合程度的评估及模型参数的统计检验上扮演着重要角色，同时也提供了一种计算残差的方法，后者可被用来探测特异值。

探测广义线性模型中的特异案例

与在OLS回归中一样，异常案例同样能扭曲GLM的估计。对于某些模型，如二分类logit模型和probit模型，特异案例的影响通常没有那么严重，因为因变量本身就只能取两个可能值，但此类特异案例仍然有可能影响回归估计。对于其他模型，如泊松模型，高度异常的因变量取值更容易出现。因此，对特异值的探测在GLM中同样重要。很多用于OLS回归的诊断技术经过调整即可

用于 GLM,其中用于探测特异观察案例的那些技术都很有效。

广义线性模型的残差(Residuals From the GLM)

广义线性模型的残差有好几种定义方式。其中一些,包括回应残差(response residual),即观察值 y 与其预测值之差 $y_i-\hat{\mu}_i$;偏差残差(deviance residual),从模型偏差的案例成分(case-wise components)衍生而来;工作残差(working residual),加权最小二乘法最少一次迭代时的残差。另外还有学生化残差的类似物(approximations of studentized residual)。本书最关心的是皮尔森残差(Pearson residual),因为它们在很多稳健 GLM 模型中扮演着核心角色。皮尔森残差就是经由期望值标准差尺度化的回应残差:

$$e_{Pearson_i}=\frac{y_i-\mu_i}{\sqrt{V(\hat{\mu})}} \qquad [6.12]$$

关于各类残差的相对优点的更多细节,可参看 Gill, 2001。这些残差都有各自的用途,但都难以同时适用于所有目标。

帽子值与杠杆效应(Hat Values and Leverage)

和 OLS 回归一样,GLM 模型中的杠杆效应也可以通过帽子值 h_i 进行衡量,它们来自于迭代过程中最后一次 IWLS 估计。不过,和在线性回归中不同,GLM 模型中的帽子值同时取决于 y 值和 x 值。遵照普雷吉本的观点(Pregibon, 1981),帽子矩阵是这样定义的:

$$H=W^{1/2}X(X^TWX)X^TW^{1/2} \qquad [6.13]$$

其中 W 为来自 IWLS 拟合最后一次迭代的权数矩阵。这一帽子矩阵因为 X 被替换为 $W^{1/2}X$ 而不同于一般形式的 H(方程 3.7)。如此处理使得 y 的方差可以有所变化,也正因如此,帽子值取决于 y 和 X 二者(参见 McCullagh and Nelder, 1989:405)。

权势的评估(Assessing Influence)

和在线性模型中一样,DFBETA 和库氏距离在探测广义线性模型(GLM)

中的权势值上大有用处。DFBETA 是在比较移除某个观察案例前后回归系数的差异的基础上计算出来的，$D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}$，$i = 1, \cdots, k$。类似于库氏距离的权势测量指标在 GLM 中同样可得：

$$D_i = \frac{e^2_{Pearson_i}}{\hat{\varphi}(k+1)} \times \frac{h_i}{1-h_i} \qquad [6.14]$$

其中 $\hat{\varphi}$ 是模型离散度估计值，而 k 为模型中除截距以外其他待估参数的个数（见 Fox，2002）

稳健广义线性模型

GLM 模型的稳健估计方法的发展远远落后于线性模型的稳健估计。虽然有几个早期的尝试试图使 logistic 回归更加稳健（如 Pregibon，1981；Copas，1988；Carroll and Pederson，1993；Bianco and Yohai，1996），但对于其他 GLM 模型的拓展仍然很少被考虑。即使在今天，也只有少数几款统计软件含有估计稳健 GLM 模型的程序，而且即使有，通常也仅限于 logit 模型和泊松模型。

GLM 模型的 M 估计(M-estimation for GLMs)

和线性模型一样，GLM 模型中使用最广泛的稳健方法在某种形式上也是基于 M 估计。与线性回归的早期 M 估计一样，GLM 模型的 M 估计的早期尝试同样面临影响函数无界这一问题（见 Stefanski，Carroll，and Ruppert，1986；Kunsch，Stefanski，and Carroll，1989）。通常，得到的估计量也不可取，因为它们具有 Fisher 不一致性(Fisher inconsistent)。[25]不过近年来，已经发展出一些基于拟似然估计的一致性有限影响方法（consistent bounded influence methods）。其中之一由康冬尼和罗切蒂提出（Cantoni and Ronchetti，2001）。[26]

康冬尼和罗切蒂的估计量由对 Preisser 和 Qaqish 的方程（1999）的拟似然广义估计推演而来：

$$\sum_{i=1}^{n} \frac{\partial}{\partial \beta} Q(y_i; \mu_i) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \mu'_i = 0 \qquad [6.15]$$

其中 $\mu_i'=\frac{\partial}{\partial\beta}\mu_i$ 和 $Q(y_i;\mu_i)$ 为拟似然函数。它的解是由下列得分函数确定的 M 估计量：

$$\Psi(y_i;\mu_i)=\frac{(y_i-\mu_i)}{v(\mu_i)}\mu_i' \qquad [6.16]$$

不过，这一估计量难以用于稳健回归，因为它的影响与 Ψ 成比例，因而没有界限。

康冬尼和罗切蒂遵照马娄斯的 GM 回归估计的逻辑对方程 6.16 进行了改进。还记得一般 M 估计量是如下方程的解：

$$\sum_{i=1}^{n}\Psi(y;\theta)=0 \qquad [6.17]$$

或者，对于广义线性模型这一具体情况：

$$\sum_{i=1}^{n}\Psi(y;\mu)=0 \qquad [6.18]$$

其中 Ψ 提供了观察案例的权数。和线性模型的 MM 估计相似，如果该函数为奇函数且有界，即意味着 $\rho(\cdot)$关于 0 对称，那么该估计量的崩溃点 BDP 就等于 0.5。康冬尼和罗切蒂是通过求解下面的方程达到这一点的：

$$\Psi(y;\mu)=v(y;\mu)w(x)\mu'-a(\beta) \qquad [6.19]$$

其中：

$$a(\beta)=\frac{1}{2}\sum_{i=1}^{n}E[v(y_i;\mu_i)]w(x_i)\mu_i \qquad [6.20]$$

而 v_i 和 w_i 为同时考虑了相应案例的残差及帽子值的权数函数。对 Huber 函数的改造和使用保证了所获权数对特异 y 值具有稳健性：

$$v_i(y_i;\mu_i)=\Psi(e_i)\frac{1}{V^{1/2}(\mu_i)} \qquad [6.21]$$

仿照马娄斯提供的线性模型 GM 估计量，$w_i(x_i)$ 的一个可能选择是 $w_i(x_i)=\sqrt{1-h_i}$。不过，我们已经看到，这会导致过低的崩溃点，因此这里使用的是稳健距离的倒数(请参见第 4 章关于稳健距离的讨论)。最终的结果是一个有效率的

估计量，具有有界影响，并且服从渐近正态分布。更重要的是，已有研究显示，当数据被污染时，从这种模型作出的推论比从一般 GLM 模型作出的推论要可靠得多(参见 Cantoni and Ronchetti，2001)。

例 6.1：预测 2001 年是否投票给英国工党的 Logistic 模型

这一例子使用的数据来自于 1997—2001 年英国选举跟踪研究(British Election Panel Study)(Heath，Jowell and Curtice，2002)。我们只关注那些参加了 2001 年最后一轮调查的受访者。在清除缺失值后，最终分析样本的规模为 1421。我们的分析目标是评估工党领袖托尼·布莱尔对受访者在 2001 年英国大选期间是否投票给工党的影响。因变量为投票给工党(编码为 1)或其他党(编码为 0)。对于布莱尔的评价采用的是一个 5 分李克特题目，询问受访者对布莱尔作为英国首相的表现的好坏评价(取值越高代表做得越好)。分析中控制了年龄、性别、教育水平(本科及以上学位，接受了一些高等教育[some postsecondary]，大学预科[a level]，高中[o level]，无)，社会阶层(管理者/专业人员，常规非体力劳动者，自雇佣者，体力劳动者)；主观认知的社会经济地位的变化(五分量表，较高的数字表示受访者认为自己的个人社会经济地位在过去 5 年得到改善)。[27] 对这一数据，我们同时拟合了常规 logistic 回归和稳健回归。

首先来评估(各案例在)常规 logit 模型中的权势。正如我们在图 6.1 的 Cook's D 标号图中看到的，有好几个观察案例对回归平面的相对权势很高。不过，通过进一步的诊断检查，包括对每一个系数的 $DFBETA_i$ 的仔细检查，并未发现任何明显的问题。换句话说，尽管一些案例整体上有着异常高的权势，但它们似乎对所有系数都不存在严重影响，至少单个看来是这样。考虑到这些案例整体的权势较高，我们仍然尝试了一下稳健回归，看它能否给出与常规 logistic 回归不同的结果。

表 6.2 给出了两个回归的结果。尽管出现了一些相对权势较高的观察案例，但常规 logistic 回归的表现仍然很好。事实上，这两个模型的主要结论极其相似——可以认为对于布莱尔的评价深刻影响了受访者是否投票给工党。尽管对于布莱尔的评价的系数在稳健 logistic 回归中稍微大些(1.205 比 1.127)，

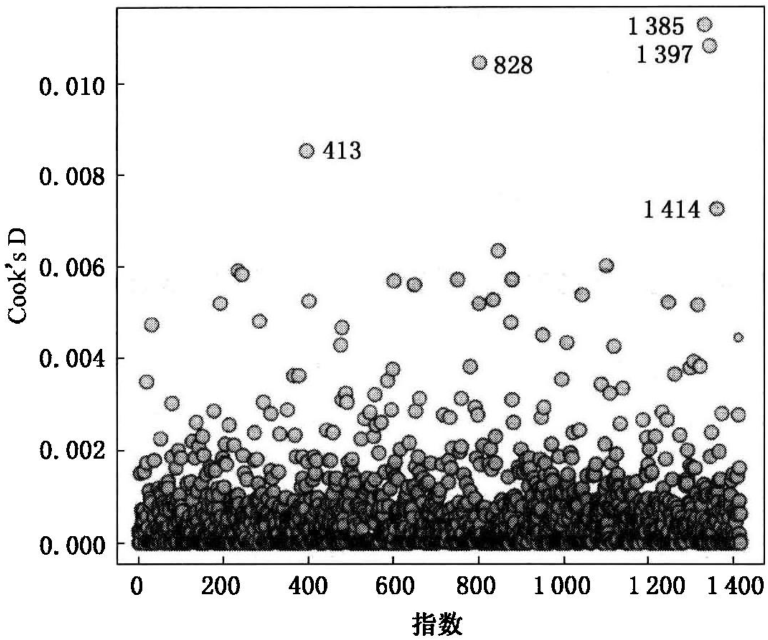


图 6.1 预测 2001 年是否投票给了英国工党的 Logistic 模型的 Cook's D 标号图

表 6.2 预测 2001 年是否投票给英国工党的 Logistic 模型

	最大似然估计 logit 模型		稳健 logit 模型	
截距	−5.15	0.468	−5.42	0.525
年龄	−0.003	0.004	−0.003	0.004
男	0.117	0.141	0.129	0.144
教育水平				
本科以上学位	−0.372	0.269	−0.350	0.276
大学预科	−0.391	0.255	−0.321	0.261
高中	−0.190	0.181	0.163	0.187
部分高等教育	−0.462	0.235	−0.374	0.241
无	0	—	0	—
社会阶层				
专业人员及管理者	−0.055	0.184	−0.040	0.189
常规非体力劳动者	−0.271	0.193	−0.213	0.197
自雇佣者	−0.548	0.259	−0.551	0.266
体力劳动者	0	—	0	—
主观认知的经济变化				
社会经济状况变化	0.496	0.087	0.476	0.090
自我经济状况变化	0.268	0.077	0.266	0.079
对布莱尔的态度	1.127	0.101	1.205	0.122
n	1421		1421	

但这两个系数之间的差异在统计上并不显著。因此，对于这一数据，我们应该优先选择常规 logistic 回归，因为它比稳健回归简单。这个例子非常典型地说明了特异观察案例很难对 logistic 回归平面产生严重的影响，因为因变量只能取两个值。不过，下面我们将看到，特异案例更容易对泊松回归产生严重影响。

例 6.2：对魁北克居民的自愿性组织成员资格进行解释的稳健泊松回归

这个例子使用的数据来自 2000 年加拿大平等、安全及社区调查(Canadian Equality, Security, and Community survey of 2000)。尽管该数据包括来自加拿大不同地区的受访者的信息，但这里仅将来自魁北克的受访者纳入分析 ($n=949$)。因变量为受访者参加的自愿性组织的数量。自变量包括性别(妇女作为参照组)，出生在加拿大(参照组为“非出生在加拿大”)，在家使用的语言(分为英语、法语及其他，法语被编码为参照组)。由于因变量为计数变量(且服从泊松分布)，所以这里使用了泊松回归。我们既拟合了使用最大似然估计的常规广义线性模型，也拟合了使用拟似然估计的稳健 GLM 模型。在讨论结果之前，我们先来看 OLS 回归的诊断图。

尽管我们进行了大量的诊断，但这里只报告了那些被发现存在潜在问题的案例。图 6.2 给出了“出生在加拿大”这一变量的库氏距离及 $DFBETA_i$ 的标号图。从库氏距离来看，大概有 10 个观察案例对回归的影响很大，其中两个尤其存在问题(案例 770 和 3773)。对于 $DFBETA_i$ 的分析表明，这两个案例对于“出生在加拿

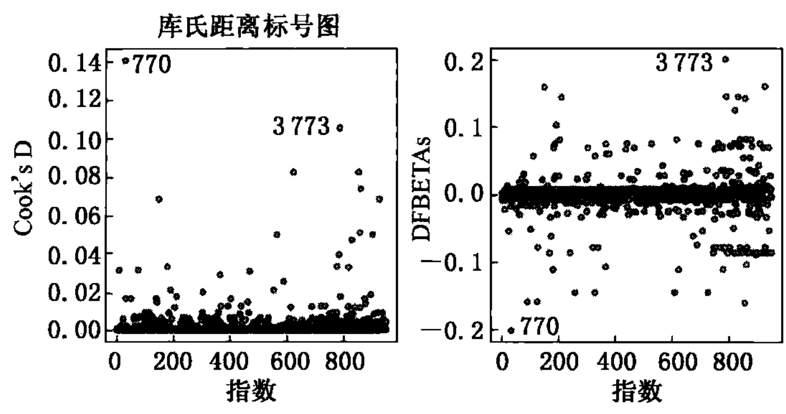


图 6.2 对魁北克居民的志愿性组织参与度进行解释的泊松模型的诊断图

大"这一变量的效果的影响非常大,尽管从标号图中可以看到它们的影响方向相反。

表 6.3 给出了来自常规泊松回归和稳健泊松回归的结果情况。可以清楚地看到,在常规 GLM 模型中,是否出生在加拿大这一变量的系数受到了与数据主体模式不符的特异观察案例的影响。稳健回归模型中这一变量的系数达到常规 GLM 模型中的 10 倍。作用(系数)上的这种差异将导致非常不同的实际解释。如果是基于常规 GLM 模型,我们将得出结论认为:在控制其他变量的情况下,出生在加拿大的人与出生在其他地方的人在志愿性组织的参与上不存在差异 ($e^{0.0027} = 1.03$; $p = 0.71$)。相反,稳健回归告诉我们,平均而言,在其他变量固定不变的情况下,出生在加拿大的人参与的组织数要比出生在其他地方的人多 30% ($e^{0.258} = 1.29$; $p = 0.0035$)。

表 6.3 对志愿性组织成员资格进行解释的泊松回归模型

	最大似然 GLM 模型			稳健 GLM 模型		
	$\hat{\beta}$	SE($\hat{\beta}$)	$e^{\hat{\beta}}$	$\hat{\beta}$	SE($\hat{\beta}$)	$e^{hat\beta}$
截距	0.586	0.077	1.79	0.120	0.095	1.13
男	0.079	0.045	1.08	0.084	0.053	1.09
生于加拿大	0.027	0.072	1.03	0.258	0.088	1.29
语言						
英语	0.357	0.061	1.43	0.537	0.068	1.71
其他	−0.014	0.094	0.98	0.079	0.112	1.08
法语	0	0	1.00	0	0	1.00
n	949			949		

本章的两个例题都大有教益,原因有二:第一,泊松回归的例子清楚地表明 GLM 的估计能够被特异观察案例严重影响。基于常规 GLM 模型的结论与基于稳健 GLM 模型的结论大不相同。因此在这个例子里,报告稳健 GLM 模型结果更合理。第二,logistic 回归例子显示,即使存在大量权势很高的特异案例,稳健 GLM 模型的实际结论也并不一定有别于常规 GLM 模型所得出的结论。因为此时的因变量只能取两个值——因此残差通常不会特别大——对 logistic 回归来说,通常总是这样。在这种情况下,常规 GLM 模型因为比稳健 GLM 模型简洁而更受欢迎。但是,作为一种诊断工具,GLM 模型仍然值得探测。

第7章 | 结论

本书特别强调探测和恰当处理回归分析中特异案例的重要性。书中的经验示例表明:如果此类案例未被察觉,它们将严重扭曲回归估计。书中还提供了一些证据表明,垂直异常值,更一般地讲,重尾分布将降低回归估计的精度。这些问题同时适用于采用OLS估计以及更为一般的GLM估计所拟合的模型,这进一步突出了回归诊断的重要性。在本书中,我们还对一些用来探测垂直异常值、杠杆点,以及权势值的传统方法进行了介绍。通过组合起来使用,这些方法在识别经验例子中的问题观察案例时非常有效。

发现问题案例以后,有几种处理方式可供研究者考虑。最简单的"处理"就是直接将这些讨厌的案例从分析中移除。如果有很好的理由这样做,比如某个案例被错误编码或者已知由于某种原因而非常独特,那么移除是很好的策略。然而,有时特异案例反映的是模型无法解释的系统性问题。这一点非常重要,因为它意味着特异案例并非总是等同于"坏"的数据。事实上,特异值很可能是数据中最具吸引力的部分。如果出现的异常观察案例很多,我们可以通过向模型添加新的因素来应对这种差异,这些因素可以是新的变量,也可以是既有变量之间的交互项。如果没有完全合理的理由支持我们对于特异案例的移除或者模型设置的改变,那么稳健回归技术就是合适的选择。

一方面,稳健回归的策略并非显著地区别于移除异常案例。两种策略的目标都是为数据的主体部分寻找拟合最优的模型。关于这一点,有人可能会批评两者都会导致数据截除(truncation)偏差。换句话说,当我们尚不清楚这些异常案例是否真的是"被污染"数据时,将其移除或者降低其权重,都会使我们的回归估计有偏。我并不认同这一说法。使用统计模型是为了描述数据的一般模

式，目的应该是从数据中讲述最具可能性的“故事”。因此，如果我们谈论的仅仅是由于一个或几个异常案例造成的 y 和 x 的关系，不管其在统计上显著与否，都是不合理的。这将是一个误导性的故事。相反，我们应当更加关注那些符合数据主体部分，而非少数特选案例的模式。当然，这并不表示我们可以不考虑特定观察案例异常的原因，就机械地报告稳健回归结果。事实上，我主张将这些方法主要作为诊断工具来使用，只有当我们确实无法解释这种异常性时，才靠它们来决定最终模型。

本书讨论了许多不同类型的稳健回归，尽管其中多数早期方法已经相当过时。介绍如此多的方法之所以必要，是因为新近方法都建立在早期方法的基础之上。考虑到我们的主要目标是克服特异案例对回归估计的影响，因此，应当选择那些崩溃点高且影响有限的估计量。许多早期的稳健回归方法，如 LMS, LTS, LTM 以及 S-估计，都符合这些标准。但这些方法在高斯—马科夫假定下效率很低。如果分析的目的在于从样本数据推论到作为其来源的更大总体，那我们还需要有效率的估计。基于 M 估计的稳健回归模型就符合这一标准。但是，原始的 M 估计并不具备很强的抗扰性。它们的崩溃点是 0，这意味着仅仅一个(特异)观察案例就能严重影响其估计效果。

幸好，广义的 M 估计将高抗扰性和高效率结合了起来。这些估计量是这样达到该属性的：首先通过具有高抗扰性的模型(如 LMS, LTS, LTM 或 S—估计)找到残差和/或其尺度的最初估计，然后将这些信息用于 M 估计后续阶段以使残差最小化。最重要的两个估计量是——至少在稳健回归被用于作出结论的最终模型时——寇克里和赫特曼斯伯格(Coakley and Hettmansperger, 1993)的 GM 估计量，即我们所知道的一步史威普估计量和由约哈依(Yohai, 1987)最先提出的 MM 估计量。前者虽然能最佳地处理杠杆点，但是它们在小样本中非常缺乏效率。相反，后者在多数条件下都表现得相当出色。不过，正如第 4 章所示，早期“抗扰性”有余但缺乏效率的估计量在稳健回归用作诊断工具时也能够起到重要的作用。

尽管崩溃点高的估计量具有明显的吸引力，但在使用时也要非常小心。一项针对这些技术的批评是，标准诊断方法在探测曲度或非线性(curvature)方面

存在问题(Cook，Hawkins and Weisberg，1992；McKean，Sheather and Hettmansperger，1993)。韦尔考克斯(Wilcox，2005)建议在使用高崩溃点的方法时，应该配合使用其他曲度探测效果更好的方法以对稳健回归进行补充。在这方面，非参数回归和广义可加模型(generalized additive model，GAM)将会很有帮助。GAM不在本书范围之内，但是我建议读者看哈斯蒂与提布希拉尼的著作(Hastie and Tibshirani，1990)，里面有关于这些模型的详尽讨论，福克斯也有好的入门介绍(Fox，2000a，2000b)。

我希望本书已清楚阐明我们不应盲目使用任何一种方法。除非数据非常好(换言之，残差服从正态并且没有特异值)，由于数据形制不同，不同回归技术会给出截然不同的答案。明智的做法是在数据分析的准备阶段，就使用稳健回归方法和OLS及相关的诊断方法。将这些方法得出的系数简单对比，往往就能得知OLS回归的有效性。如果估计结果迥异，用稳健回归残差相对于OLS回归残差绘制而成的RR图，能够显示出是哪些观察案例引发了差异。将这些诊断方法与传统方法，如库氏距离和DFBETA等相结合，将会很有帮助。最后，OLS回归的相对效率和简易性表明，如果它确实能为数据中的模式提供相当好的刻画，那么选择使用它才是明智的选择。这些原则同样适用于(常规)GLM回归和对应的更稳健的替代方法。

附录 | 稳健回归的软件选择

本书呈现的所有统计分析都是用 R 软件(R Development Core Team, 2006)完成的,它执行的是 S 语言。除了卓越的功能性和灵活性之外,R 软件也因其免费可得而具有很强的吸引力。你可以在 http://cran.r-project.org 下载 R 软件的各种操作系统的版本。除了基本 R 软件配有的标准组件(或数据包)和推荐组件外,它还有很多附加数据包(在 S 语言中称为"库"),这些也是免费的。通过快捷的互联网,我们能方便地获取并在 R 软件内安装这些组件。

尽管本书使用的很多方法都可以在若干数据包中找到,但下面列出的这些非常有用。"MASS"包(Venables and Ripley, 2002)包括 M 估计、S 估计及 MM 估计函数。"robustbase"包(Maronna et al., 2006)拥有各种不同稳健回归方法的函数——包括 MM 估计,LTS 估计,以及稳健广义线性模型——和基于稳健距离基础上的一些稳健诊断图。在其他数据包中,"car"包(Fox, 2002)包括线性模型和广义线性模型的传统诊断函数。"boot"包(Davidson and Hinkley, 1997,但由 Angelo Canty 撰写)提供了出色的一般自助函数,它们被用来产生前面报告的稳健回归自助标准误。"quantreg"包(Koenker, 2005)包含用于拟合 LAV 回归的函数。最后,由韦尔考克斯(Wilcox, 2005)编写的免费 R 代码囊括了许多有用的稳健程序(robust routines),包括本书提到的 GM 估计量。本书全部案例所用数据和 R 代码都可通过本书的网址下载:www.sagepub.com/andersendata。

上面讨论的所有数据包在 SPlus 中也都能找到。SPlus 中最全面而丰富的稳健回归数据包是"robust"库,它可以用来拟合线性模型的 M 估计、MM 估计以及其他稳健估计。它同样可以用来拟合各种稳健的 GLM 模型,包括 Kunsch

等人(1989)的模型。"robust"包对 R 软件同样可用,但不同于其他上面提及的 R 包,使用它要向 Insightful(SPlus 的制造商)的支付许可费。

SAS 和 Stata 也有针对稳健回归的良好功能,尤其是各种 M 估计量和 L 估计量。SAS 的 PROGRESS 程序拥有进行 LMS 和 LTS 回归的程序,并且 ROBUSTREG 程序(在第 9 版中最先出现)能够实现常用稳健回归技术的绝大多数,包括 M 估计,LTS, S 估计和 MM 估计。在 Stata 中,*rreg* 命令能够实现基于 M 估计的各种稳健回归,包括 MM 估计。而 *qreg* 命令能够用于 LAD 回归和 LAV 回归。自助标准误在 Stata 中则可以通过 *bsreg* 函数轻松计算得到。目前,稳健 GLM 回归在 SAS 和 Stata 中都还无法实现。

Gauss 软件也提供了各种稳健回归选择,包括 LAD(和分位回归,更概括地说),以及各种类型的 M 估计。在拟合稳健回归时,程序会自动提供自助标准误。尽管在稳健回归性能的扩展方面较为欠缺,LIMDEP 软件同样可以用来拟合 LAD, LAV 回归,以及一些 M 估计,并为这些估计提供自助标准误。任何一种稳健回归程序——不论是用于线性模型还是 GLM 模型——在 SPSS 中都不存在。

注释

[1] 通常是用 ϵ_n^* 而非 BDP 来表示崩溃点。这里我特意不使用 ϵ_n^*，以避免它与回归模型中的误差项相混淆，后者与崩溃点无关。

[2] 尽管关于崩溃点的讨论经常使用偏差(*bias*)的概念，但这里使用影响(*effects*)是为了避免与此前已经讨论过的偏差的一般统计学意义相混淆。如果影响很大的特异值并不是错误编码的结果，那么不管特异值对估计的影响多大，一个估计量仍然可以是无偏的——也就是说，重复随机抽样产生的该估计量的平均值将等于总体的参数。不过，这并不意味着该估计量是这个数据的一个好描述。

[3] 如果估计量满足这一条件，即被认为达到了 Cramer-Rao 下限(具体细节见 Cramer，1946)。

[4] 比克尔和莱曼(Bickel and Lehmann，1975)还建议了第 5 个条件。假定两个随机变量 X 和 Y。如果 X 随机地大于 Y——也就是说，对于任意 x 值，$F_X(x) \leqslant F_Y(x)$ ——那么作为位置测度，必须满足 $\theta(X) \geqslant \theta(Y)$。

[5] 尽管这是最常见的对截尾均值的定义，也曾有人提出过略有差异的估计量(见 Reed，1998；Kim，1992)。

[6] 休伯(Huber，2004)进一步指出，截尾均值的影响函数即使在剪除不对称时也能求导出来。

[7] 离散程度的度量是一种特殊的尺度测度，当 X 和 Y 都是对称分布，而 $|y|$ 的随机分布大于 $|x|$ 的随机分布时，$\tau(X) \geqslant \tau(Y)$ (Kickel and Lehman，1976)。

[8] MAD 有时被错误地用来指称不太稳健的"相对均值的中位离差"。离均中位离差用得很少，因为当使用它的条件——单峰对称分布——满足时，标准差有用得多。

[9] 从第 6 章可以看到，在广义线性模型中并不是这种情况。

[10] 不过，在样本规模很大时，这一分界线难以确认任何观察案例，不管它们是否值得注意(Fox 1991)。

[11] 福克斯的权势图通过使用 R 组件包(car package for R)中的 influence. plot 功能很容易就能画出来(更具体的内容请看 Fox，2002:198)。

[12] 不要把偏回归图混淆为相似的偏残差图。后者在评估权势上没那么有效，但在区分单调和非单调的非线性上更胜一筹。关于偏回归图和偏残差图的相对优劣的更多内容，可以看福克斯的作品(Fox，1997)。

[13] LAV 其他的名字还有最小绝对离差(LAD)回归和最小绝对误差和(MSAE)回归(Birkes and Dodge，1993)。

[14] 因为使用有限而没有在本书中讨论的相关方法包括最小截尾中位数估计量(least-trimmed median estimators)和最小截尾差异估计量(least-trimmed difference estimator)。这两种估计量的崩溃点 $BDP = 0.5$，但它们的相对效率小于 67%。更多内容，可以看克鲁克斯或斯特龙伯格的作品(Croux et al.，1994；Stromberg et al.，2000)。

[15] 这些估计量有时也被叫做截尾均值估计量(trimmed-mean estimator)。它们也能被调整从而具有有限影响函数(见 De Jongh，De Wet，and Welsh 1988)。

[16] 不要将 LMS 估计量混淆为西格尔(Siegel，1982)的重复中位数(RM)。尽管很早以前就被当作稳健估计量，但 RM 估计量存在严重不足，它在高维问题下不具仿射回归同变性(not being affine-regression equivariant)。也就是说，当自变量被重新尺度化或线性组合时，系数估计不能如预期的那样表现。由于这一限制，本文将不再进一步讨论。

[17] IRLS 也被称为迭代加权最小二乘法(iterative weighted least squares, IWLS)。

[18] 这里假定 Huber 权数的微调常量为 $c=1.345$,双平方权数的为 $c=4.685$。更多细节请看前面关于位置的 M 估计的讨论。

[19] 像 LMS 估计(Rousseeuw, 1984)和 RM 估计(Siegel, 1982)等其他方法也曾被提出来用于初始估计。

[20] 前面已经说过,稳健回归的标准误有时完全不同于更常使用的"稳健标准误(robust standard errors)",后者常被用来校正未知的异方差模式。稳健标准误有很多熟知的名称,如 White 标准误、Eicker 标准误和 Huber 标准误,和其他更为一般的"三明治估计量(sandwich estimators)"一样,它们的运算过程中并没有对 OLS 回归本身做任何改变。如果想了解更多有关稳健标准误的一般知识可以看 White(1980)。关于几种稳健标准误的表现的出色讨论,可以看 Long and Ervin (2000)。

[21] 有关使用似然法计算标准误的描述可以看 Western(1995),虽然简短但非常出色。

[22] 如果样本规模太小,自助法也会出现故障。Chernick(1999:151)建议样本规模至少应该达到 30。另外,自助法也可能因为相依数据(dependent data),或因为数据中存在相当比例的非随机缺失值而出现故障(见 Davidson and Hinkley, 1997:37—54; Chernick, 1999:102—105)。

[23] 如果样本规模很大,对线性模型采取自助法得到的标准误将和常规标准误接近。

[24] 折刀法再抽样的方式不同于自助法,它不是从数据中回置性地抽取随机样本,而是通过随机地移除一个案例来进行再抽样(通常进行 n 次)(更多细节可以看 Davidson and Hinkley, 1997:113—118)。

[25] 一个 M 估计如果满足如下条件就被认为是条件性地 Fisher 一致(conditionally Fisher-consistent):

对于所有 β 和 x 都有 $E_\beta[\Psi(y, x, \beta) \mid x] = \iint \Psi(y, x, \beta) P_\beta(dy \mid x) = 0$

如果 x 的分布独立于 β,那么线性模型和广义线性模型的最大似然估计量就属于条件性的 Fisher 一致。

[26] R 软件 robustbase 库中的 glmrob 函数用的就是这种方法。

[27] 有关变量编码的详细信息参见 Aderson 和 Evans(2003)。

参考文献

Andersen, R. and G. Evans. 2003. "Who Blairs Wins? Leadership and Voting in the 2001 Election." *British Elections & Parties Review*, 13:229—47.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. 1972. *Robust Estimates of Location*. Princeton, NJ: Princeton University Press.

Atkinson, A. C. 1985. *Plots, Transformations and Regression*. Oxford, England: Clarendon.

Atkinson, A. and M. Riani. 2000. *Robust Diagnostic Regression Analysis*. New York: Springer-Verlag.

Belsley, D. A., E. Kuhn, and R. E. Welsch. 1980. *Regression Diagnostics*. New York: Wiley.

Bianco, A. M. and V. J. Yohai. 1996. "Robust Estimation in the Logistic Regression Model." Pp. 17—34 in *Robust Statistics, Data Analysis, and Computer Intensive Methods*, edited by H. Rieder. New York: Springer-Verlag.

Bickel, P. J. and E. L. Lehmann. 1975. "Descriptive Statistics for Nonparametric Models Ⅱ. Location." *Annals of Statistics* 3:1045—69.

Bickel, P. J. and E. L. Lehmann. 1976. "Descriptive Statistics for Nonparametric Models Ⅲ. Dispersion." *Annals of Statistics* 4:1139—58.

Birch, J. B. and D. B. Agard. 1993. "Robust Inferences in Regression: A Comparative Study." *Communications in Statistics, Simulation and Computation* 22:217—44.

Birkes, D. and Y. Dodge. 1993. *Alternative Methods of Regression*. New York: Wiley.

Cantoni, E. and E. Ronchetti. 2001. "Robust Inference for Generalized Linear Models." *Journal of the American Statistical Association* 96: 1022—30.

Carroll, R. J. and S. Pederson. 1993. "On Robustness in the Logistic Regression Model." *Journal of the Royal Statistical Society, Series B* 55:693—706.

Carroll, R. J. and A. H. Welsh. 1988. "A Note on Asymmetry and Robustness in Linear Regression." *American Statistician* 42:285—87.

Chatterjee, S. and A. S. Hadi. 1988. *Sensitivity Analysis in Linear Regression*. New York: Wiley.

Chave, A. D. and D. J. Thomson. 2003. "A Bounded Influence Regression Estimator Based on the Statistics of the Hat Matrix." *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 52:307—22.

Chernick, M. R. 1999. *Bootstrap Methods: A Practitioner's Guide*. New York: Wiley.

Coakley, C. W. and T. P. Hettamansperger. 1993. "A Bounded Influence, High Breakdown, Efficient Regression Estimator." *Journal of the American Statistical Association* 88:872—80.

Cook, R. D. 1977. "Detection of Influential Observations in Linear Regression." *Technometrics* 19: 15—18.

Cook, R. D. and D. M. Hawkins. 1990. "Unmasking Multivariate Outliers and Leverage Points: Comment." *Journal of the American Statistical Association* 85:640—44.

Cook, R. D. and S. Weisberg. 1982. *Residuals and Influence in Regression*. London: Chapman Hall.

Cook R. D. and S. Weisberg. 1999. *Applied Regression including Computing and Graphics*. New York: Wiley.

Cook, R. D., D. M. Hawkins, and S. Weisberg. 1992. "Comparison of Model Misspecification Diagnostics Using Residuals From Least Mean of Squares and Least Median of Squares Fit." *Journal of the American Statistician* 87:419—24.

Copas, J. B. 1988. "Binary Regression Models for Contaminated Data." *Journal of the Royal Statistical Society, Series B* 50:225—65.

Cramer, H. 1946. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Croux, C., P. J. Rousseeuw, and O. Hossjer. 1994. "Generalized S-Estimators." *Journal of the American Statistical Association* 89:1271—81.

Davidson, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge, England: Cambridge University Press.

Davis. J. B. and J. W. McKean. 1993. "Rank-Based Methods for Multivariate Linear Models." *Journal of the American Statistical Association* 88:245—51.

De Jongh, P. J., T. De Wet, and A. H. Welsh. 1988. "Mallows-Type Bounded-Influence-Regression Trimmed Means." *Journal of the American Statistical Association* 83:805—10.

Dietz, T. R., S. Frey, and L. Karloff. 1987. "Estimation With Cross-National Data: Robust and Nonparametric Methods." *American Sociological Review* 52:380—90.

Dobson, A. J. 1990. *An introduction to Generalized Linear Models*. New York: Chapman Hall.

Draper, N. R. and H. Smith. 1998. *Applied Regression Analysis*. New York: Wiley.

Dunteman, G. H. and M. R. Ho. 2005. *An Introduction to Generalized Linear Models* (Quantitative Applications in the Social Sciences, 07—145). Thousand Oaks, CA: Sage.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7:1—26.

Efron, B. 1981. Nonparametric Standard Errors and Confidence Intervals(With Discussion). " *Canadian Journal of Statistics* 9:139—72.

Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman Hall.

Fahrmeir, L. and G. Tutz. 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2d ed. New York: Springer.

Ferretti, N., D. Kelmansy, V. J. Yohai, and R. H. Zamar. 1999. "A Class of Locally and Globally Robust Regression Estimates." *Journal of the American Statistical Association* 94:174—88.

Fox, J. 1991. *Regression Diagnostics: An Introduction* (Quantitative Applications in the Social Sciences, 07—079). Newbury Park, CA: Sage.

Fox, J. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.

Fox, J. 2000a. *Simple Nonparametric Regression*. (Quantitative Applications in the Social Sciences, 07—129). Thousand Oaks, CA: Sage.

Fox, J. 2000b. *Multiple and Generalized Nonparametric Regression*. (Quantitative Applications in

the Social Sciences, 07—130). Thousand Oaks, CA: Sage.

Fox, J. 2002. *An R and S-PLUS Companion to Applied Regression*: Thousand Oaks, CA: Sage.

Fung, W. K. 1999. "Outlier Diagnostics in Several Multivariate Samples." *Statistician* 48:73—84.

Gill, J. 2001. *Generalized Linear Models. A Unified Approach* (Quantitative Applications in the Social Sciences, 07—134). Thousand Oaks, CA: Sage.

Hampel, F. R. 1974. "The Influence Curve and Its Role in Robust Estimation." *Journal of the American Statistical Association* 69:383—93.

Hampel, F. R. 1975. "Beyond Location Parameters: Robust Concepts and Methods." *International Statistical Institute, Proceedings of the 40th Session* 46:375—91.

Hampel, F. R. E. Z. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Handschin, E., F. C. Schweppe, J. Kohlas, and A. Fiechter. 1975. "Bad Data Analysis for Power System State Estimation." *IEEE Transactions of Power Apparatus and Systems, PAS-94*, 329—37.

Hao, L. and D. Q. Naiman. 2007. *Quantile Regression* (Quantitative Applications in the Social Sciences, 07—149). Thousand Oaks, CA: Sage.

Hastie, T. J. and R. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman Hall.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

Heath, A., R. Jowell, and J. K. Curtice. 2002. *British Election Panel Study, 1997—2001: Waves 1 to 8* [*computer file*]. 4th ed. Colchester, Essex, England: UK Data Archive [distributor], July 2002. SN: 4028.

Hill, R. W. and P. W. Holland. 1977. "Two Robust Alternatives to Least Squares Regression." *Journal of the American Statistical Association* 72:828—33.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey. 1983. *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Hogg, R. V. 1974. "Adaptive Robust Procedures." *Journal of the American Statistical Association* 69:909—27.

Huber, P. J. 1964. "Robust Estimation of Location Parameters." *Annals of Mathematical Statistics* 35:73—101.

Huber, P. J. 1973. "Robust Regression: Asymptotics, Conjectures and Monte Carlo." *Annals of Statistics* 1:799—821.

Huber, P. J. 2004. *Robust Statistics*. Hoboken, NJ: Wiley.

Inglehart, R. et al. 2000. *World Values Surveys and European Values Surveys, 1981—84, 1990—93, and 1995—97* [Data file and codebooks]. Ann Arbor, MI: ICPSR.

Jaeckel, L. A. 1972. "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals." *Annals of Mathematical Statistics* 43: 1449—58.

Jasso, G. 1985. "Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences." *American Sociological Review* 50:224—41.

Jasso, G. 1986. "Is It Outlier Deletion or Is It Sample Truncation? Notes on Science and Sexuality." *American Sociological Review* 51:738—42.

Jurečková, J. and J. Picek. 2006. *Robust Statistical Methods With R*. New York: Chapman Hall.

Jurečková, J. and P. K. Sen. 1996. *Robust Statistical Procedures: Asymptotics and Interrelations*. New York: Wiley.

Kahn, J. R. and J. R. Udry. 1986. "Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions." *American Sociological Review* 51:734—37.

Kempthorne, P. J. and M. B. Mendel. 1990. "Unmasking Multivariate Outliers and Leverage Points: Comment." *Journal of the American Statistical Association* 85:647—48.

Kenney, J. F. and E. S. Keeping. 1962. *Mathematics of Statistics*. 3d ed. Princeton, NJ: Van Nostrand.

Kim, S. J. 1992. "The Metrically Trimmed Mean as a Robust Estimator of Location." *Annals of Statistics* 26:1534—47.

Koenker, R. 2005. *Quantile Regression*. Cambridge, England: Cambridge University Press.

Koenker, R. W. and Bassett, G. W. 1978. "Regression Quantiles." *Econometrica* 46:33—50.

Koenker, R. W. and V. d'Orey. 1994. "Computing Regression Quantiles." *Applied Statistics* 43: 410—14.

Krasker, W. S. and R. E. Welsch. 1982. "Efficient Bounded-Influence Regression Estimation." *Journal of the American Statistical Association* 77:595—604.

Kunsch, H. R., L. A. Stefanski, and R. J. Carroll. 1989. "Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, With Applications to Generalized Linear Models." *Journal of the American Statistical Association* 84:460—66.

Lawrence, K. D. and J. L. Arthur, eds. 1990. *Robust Regression: Analysis and Applications*. New York: Marcel Dekker.

Lax, D. A. 1985. "Robust Estimators of Scale: Finite Sample Performance in Long-Tailed Symmetric Distributions." *Journal of the American Statistical Association* 80:736—41.

Leger, C. and J. P. Romano. 1990. "Bootstrap Adaptive Estimation: The Trimmed-Mean Example." *Canadian Journal of Statistics* 18:297—314.

Li, D. 1985. "Robust Regression." In *Exploring Data Tables, Trends and Shapes*, edited by D. Hoaglin, F. Mosteller, and J. Tukey. New York: Wiley.

Liao, T. F. 1994. *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models* (Quantitative Applications in the Social Sciences, 07—101). Thousand Oaks, CA: Sage.

Lindsey, J. K 1997. *Applying Generalized Linear Models*. New York: Springer-Verlag.

Long, J. S. and L. H. Ervin. 2000. "Using Heteroskedasticity Consistent Standard Errors in the Linear Regression Model." *American Statistician* 54:217—24.

Maronna, R. A., O. H. Butos, and V. J. Yohai. 1979. "Bias and Efficiency Robustness of General M-Estimators for Regression With Random Carriers." Pp. 91—116 in *Smoothing Techniques for Curve Estimation*, edited by T. Gasser and M. Rosenblatt. New York: Springer-Verlag.

Maronna, R. A., D. R. Martin, and V. J. Yohai. 2006. *Robust Statistics: Theory and Methods*.

New York: Wiley.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. New York: Chapman Hall.

McKean, J. W, and S. J. Sheather. 2000. "Partial Residual Plots Based on Robust Fits." *Technometrics* 42:249—61.

McKean, J. W., S. J. Sheather, and T. P. Hettmansperger. 1993. "The Use and Interpretation of Residuals Based on Robust Regression." *Journal of the American Statistical Association* 88:1254—63.

McKean, J. W. and T. J. Vidmar. 1994. "A Comparison of Two Rank-Based Methods for the Analysis of Linear Models." *American Statistician* 48:220—29.

Mooney, C. Z. and R. D. Duval. 1993. *Bootstrapping. A Nonparametric Approach to Statistical Inference* (Quantitative Applications in the Social Sciences, 07—095). Newbury Park, CA: Sage.

Mosteller, F. and J. W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison-Wesley.

Myers, R. 1990. *Classical and Modern Regression With Applications*. 2nd ed. Boston: Duxbury.

Naranjo, J. D. and T. P. Hettmansperger. 1994. "Bounded Influence Rank Regression." *Journal of the Royal Statistical Society, Series B (Methodological)* 56:209—20.

Nelder, J. A. and R. W. M. Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A* 135:370—84.

Newey, W. K. and K. D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Economerica* 55:703—8.

Pena, D. and V. Yohai. 1999. "A Past Procedure for Outlier Diagnostics in Large Regression Problems." *Journal of the American Statistical Association* 94:434—45.

Pregibon, D. 1981. "Logistic Regression Diagnostics." *Annals of Statistics* 89:705—24.

Preisser, J. S. and B. F. Qaqish. 1999. "Robust Regression for Clustered Data With Applications to Binary Regression." *Biometrics* 55:574—79.

R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsay, J. O. 1977. "A Comparative Study of Several Robust Estimates of Slope, Intercept, and Scale in Linear Regression." *Journal of the American Statistical Association* 72:608—15.

Reed, J. F. 1998. "Contributions to Adaptive Estimation." *Journal of Applied Statistics* 25:651—69.

Rousseeuw, P. J. 1984. "Least Median of Squares Regression." *Journal of the American Statistical Association* 79:871—80.

Rousseeuw, P. J. 1985. "Multivariate Estimation With High Breakdown Point." Pp. 283—297 in *Mathematical Statistics and Applications, Vol. B*, edited by W. Grossman, G. Pflug, I. Vince, and W. Wetz. Dordrecht: Reidel.

Rousseeuw, P. J. and C. Croux. 1993. "Alternatives to the Median Absolute Deviation." *Journal*

of the American Statistical Association 88:1273—83.

Rousseeuw, P. J. and A. M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, P. J. and W. C. van Zomeren. 1990. "Unmasking Multivariate Outliers and Leverage Points." *Journal of the American Statistical Association* 85:633—39.

Rousseeuw, P. J. and V. Yohai. 1984. "Robust Regression by Means of S-Estimators." *Nonlinear Time Series Analysis: Lecture Notes in Statistics* 26:256—72.

Ruppert, D. and D. G. Simpson. 1990. "Unmasking Multivariate Outliers and Leverage Points: Comment." *Journal of the American Statistical Association* 85:644—46.

Ryan, T. P. 1997. *Modern Regression Methods*. New York: Wiley.

Siegel, A. F. 1982. "Robust Regression Using Repeated Medians." *Biometrika* 69:242—44.

Stefanski, L. A., R. J. Carroll, and D. Ruppert. 1986. "Optimally Bounded Score Functions for Generalized Linear Models With Applications to Logistic Regression." *Biometrika* 73:413—24.

Stromberg, A. J. 1993. "Computation of High Breakdown Nonlinear Regression Parameters." *Journal of the American Statistical Association* 88:237—44.

Stromberg, A. J., O. Hossjer, and D. M. Hawkins. 2000. "The Least Trimmed Differences Regression Estimator and Alternatives." *Journal of the American Statistical Association* 95: 853—64.

Tukey, J. W. 1991. "Graphical Displays for Alternative Regression Fits." Pp. 309—26 in *Directions in Robust Statistics and Diagnostics*, *Part 2*, edited by W. Stahel and S. Weisberg. New York: Springer-Verlag.

Venables, W. N. and B. D. Ripley. 2002. *Modern Applied Statistics With S*. 4th ed. New York: Springer.

Weakliem, D. L., R. Andersen, and A. F. Heath. 2005. "By Popular Demand: The Effect of Public Opinion on Income Inequality." *Comparative Sociology* 4:261—84.

Wedderburn, R. W. M. 1974. "Quasilikelihood Functions, Generalized Linear Models and the Gauss-Newton Method." *Biometrika* 61:439—47.

Western, B. 1995. "Concepts and Suggestions for Robust Regression Analysis." *American Journal of Political Science* 39:786—817.

White, H. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48:817—38.

Wilcox, R. R. 2005. *Introduction to Robust Estimation and Hypothesis Testing*. New York: Elsevier Academic Press.

Wu, L. L. 1985. "Robust M-Estimation of Location and Regression." In *Sociological Methodology 1985*, edited by N. Tuma. San Francisco: Jossey-Bass.

Yohai, V. J. 1987. "High Breakdown Point and High Efficiency Robust Estimates for Regression." *Annals of Statistics* 15:642—56.

Yohai, V. J. and R. H. Zamar. 1988. "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale." *Journal of the American Statistical Association* 83: 406—13.

译名对照表

absolute deviation	绝对离差
absolute values of residuals	残差绝对值
adjusted dependent variables	调整因变量
α-trimmed mean	α 截尾均值
asymptotic standard errors	渐近标准误
breakdown points(BDP)	崩溃点
best linear unbiased estimators(BLUE)	最佳线性无偏估计量
bias corrected(BC) confidence intervals	误差修正置信区间
bisquare weights	双平方加权法
bootstrapped standard errors	自举标准误
bounded influence function	有界影响函数
bounded normal scores	有限正态得分
GM-estimators	广义 M 估计量
iteratively reweighted least squares(IRLS)	迭代再加权最小平方法
M-estimators	最大似然估计量
canonical parameters	标准参数
conditional mean	条件均值
cook's distances	库氏距离
deviance residuals	偏差残差
DFBETA	回归系数差异量
dispersion parameters	离散参数
best linear unbiased estimators(BLUE)	最佳线性无偏估计量
biweight	双权数
breakdown point(BDP)	崩溃点
simulated data and	仿真数据
trimmed-mean	截尾均值
fixed-x bootstrapping	固定 x 自举法
generalized additive models(GAM)	广义可加模型
generalized linear models(GLMs)	广义线性模型
quasi-likelihood estimation	拟似然估计
Gini coefficient	基尼系数
generalized linear models	广义线性模型
huber estimates	休伯估计
infinitesimal perturbations	极微小扰动
influence	权势

interquartile range(IQR)	四分位差
iteratively reweighted least squares(IRLS)	迭代再加权最小二乘法
jackknife resampling	折刀法再抽样
least absolute deviations(LAD) regression	最小绝对离差回归
least absolute values(LAV) regression	最小绝对值回归
least absolute deviations(LAD) regression	最小绝对离差回归
least absolute value(LAV) regression	最小绝对值回归
least median of squares(LMS) regression	最小二乘中位数回归
least quartile difference(LQD) estimator	最小四分位差估计量
least trimmed squares(LTS) regression	最小截尾二乘回归
leverage	杠杆效应
mean squared error(MSE)	均方误
partial regression plots	偏回归图
pearson residuals	皮尔森残差
quadratic weight matrix	二次权数矩阵
quasi-likelihood estimation	拟似然估计
random-x bootstrapping	随机 x 自举法
ranked residuals	秩化残差
relative efficiency	相对效率
repeated median(RM)	重复中位数
resampling	再抽样
resistance	抗扰性/耐抗性
RR-plots	残差-残差图
saturated models	饱和模型
scale equivariance	尺度同变性
studentized residuals	学生化残差
univariate outliers	单变量特异值
variance-covariance matrix	方差-协方差矩阵
vertical outliers	垂直特异值

删截、选择性样本及截断数据的回归模型

作者简介

理查德·布林

理查德·布林是北爱尔兰贝尔法斯特皇后大学社会研究中心的一名社会学教授，同时也是该中心的主席。其论文主要关注分层、社会流动、统计方法及爱尔兰社会。最近的著作是与大卫·罗德曼(David B. Rottman)合著的《阶级分层：一个比较的视角》。

译者简介

郑冰岛

郑冰岛，中国人民大学社会学系法学学士、硕士毕业。2010 年获香港科技大学社会科学部硕士学位。现为香港科技大学社会科学部博士研究生。

序

在非实验社会科学研究中，回归分析是最常用的方法。在数据收集和录入以后，研究者无一例外地开始尝试回归模型，对其定义的等式使用最小二乘法（OLS）进行估计。但 OLS 这一强大的工具却并不总是正确的。其一便是某类特殊形式的数据可能导致 OLS 估计量的偏误。布林（Breen）教授在本书中讨论的数据形式包括删截（censored）数据、选择性样本（sample-selected）数据以及截断（truncated）数据。麻烦的是，该领域的术语运用并不统一，但相信本书的例子会帮助我们澄清这些概念。

假设城市政策学者芭芭拉·布朗（Barbara Brown）希望研究这一问题：为何美国城市比其他城市在空气污染控制上花费更多？她以 Y 表示其因变量污染控制开支，并以 X_1 到 X_{10} 表示各城市从预算到社会经济指标的十项解释变量，然后从标准城市年鉴中搜集数据。设想第一种情况：在其城市样本中，年度污染开支只在超过 10 万美元时才被记录在案，否则就是缺失值。即 Y 是截断的。然而由于 X 并未被截断，而是包含所有城市的信息，因而构成删截样本。若布朗博士仍然使用 OLS 方法去估计模型，则结果如何呢？为构成数据集，她只能使用 $Y>$ 10 万美元的个案，或者她可以对所有无记录的城市假设一个小于 10 万美元的取值，如 9 万美元。无论怎样处理，OLS 都会提供有偏的参数估计。

在上面的例子中，数据的删截性（censoring）是由于因变量 Y 的截断（truncation）。而另一类更复杂的截断则是由于因变量 Y 的观测受另一变量 Z 影响。我们稍微改动空气污染的例子，以设想第二种情况：假设其他一切条件不变，但年鉴却只包含通过了空气清洁法令的城市。则变量 Z 在通过空气清洁法令时取值 1，未通过则取值为 0。这即为选择性样本问题。布林教授考虑两个步骤以回应该问题：首先，某城市通过空气清洁法令的概率有多大；其次，在通过空气清洁法令的前提下，城市的污染开支为多少，那么该模型的参数要怎样估计呢，如果不是用 OLS 模型，那么是应该使用 Tobit 模型，还是赫克曼两步骤方法，还是最大似然估计方法呢。布林教授分别对这些估计方法的弱点和优点进行讨论，如以删截回归为例，他解释了最大似然 Tobit 估计一般来说优于赫克曼两步骤方法的原因。

如布林教授所言，删截数据、选择性样本数据以及截断数据涉及“社会科学中的广泛议题”，而詹姆斯·托宾(James Tobin) 1958年的论文引发了对这类议题的现代研究。因此我们的丛书非常需要这样一本关于删截数据的著作。其次，本书也是对丛书中更早的一本《事件史分析》的有效补充，后者主要处理另一种类型的删截数据。

迈克尔·刘易斯·贝克

第 1 章 | 概论

请考虑如下问题。某次校级考试的及格成绩为 40%，且所有参加考试的学生皆被授予证书，但只有及格的学生才会同时获知考试成绩。某位研究考试成绩之影响因素的社会学家抽出一部分学生样本，试图考察一系列解释变量诸如阶级、性别、父母教育程度对学生考试成绩的影响。但其关于学生考试成绩的信息来自学生自己的考试证书。因此若以 y_i 表示第 i 位学生的考试成绩，则仅当 $y_i > 39$ 时，研究者才会得知学生的具体分数。否则(对于那些考试未及格的学生)，研究者仅仅知道 $y_i \leqslant 39$。因而研究者面临这样的问题：如何使用这种样本数据去估计考试成绩和解释变量之间的关系？有两种简单的办法。一是使用最小二乘法(OLS)对 y 进行所有解释变量的回归，该方法使用所有样本，并且对所有不及格的学生指定其 $y = 39$[1]。这种方法有许多不妥之处，而其中最重要的是 OLS 的回归系数(应该告诉我们 y 和解释变量之间的关系)显然是总体真值的偏误估计。

第二种解决办法是仅仅使用 $y > 39$ 的样本信息对 y 进行 OLS 回归。但这种方法不仅舍弃了 $y \leqslant 39$ 的所有样本信息，而且由于其估计来源于一个并不是随机选择的子样本，因而不能很好地代表总体。此处的 OLS 估计同样是总体参数的偏误估计。虽然直观上并不是显而易见的，但更重要的是，OLS 回归系数甚至也不是 $y > 39$ 的部分总体的无偏误估计(第 2 章将作解释)。

删截、选择性样本和截断数据

为了解决这一问题(这也是本书要讨论的方法)，我们需要采取两个步骤。

首先是测量个体通过考试的概率。换言之,我们使用一系列相关的解释变量来拟合 y 大于 39 的概率,即 $pr(y>39)$。然后我们再使用一列相关变量,拟合通过者的期望成绩,即 $E(y\mid y>39)$,其中 E 代表期望值。在模型拟合中,这两个步骤可以分开进行,也可更有效率地共同进行。

我们描述的此例在统计学文献中被称作删截样本问题。我们可以引入一些名称来更准确地说明其含义。若对于随机变量 y 有某数值 c,对于 $y>c$ 的所有样本,我们知道 y 的确切数值,但对于其他样本,我们则仅仅知道 $y\leqslant c$,则称为由下截断(左截断)。这正是我们开始时使用的例子所描述的情况。同时我们还有由上截断(右截断),表示我们知道所有 y 小于某一域值 c 时 y 的确切值,但对于所有其他样本,我们仅知道 $y\geqslant c$。收入是一个典型的例子,对于样本中的高收入群体,我们可能仅仅知道其年收入是 10 万美元或以上。若存在两个或更多域值,则还有可能出现多截断的情况。如两个域值 $d>c$,若 $c<y<d$,则已知 y 的具体数值;而当 $y\leqslant c$ 时,$y\leqslant c$ 即为全部已知信息;而对 $y\geqslant d$,我们仅知 $y\geqslant d$。例如高收入和低收入都被截断的例子。

假设我们有一个截断 y 的样本,其中包含一系列变量 x_k,$k=1,\cdots K$,而 y 是 x_k 的函数。则 x_k(简写为 x)是以 y 为因变量的回归分析中的解释变量。若对所有样本我们都有 x 的观察值,则样本称作删截的。所以在左删截的样本里,我们既能获得所有 $y>c$ 的 x 值(其中 y 有确切值),也可知道 y 小于或等于 c 时的 x 值。相反,如果仅仅对那些 y 有确切值的样本,其 x 才被观察到,则该样本称作截断的。在这种情况下,对于 y 缺乏具体取值的样本,我们没有任何信息。

现在我们对截断的随机变量,以及含有这类变量的整体样本数据进行区分。后者可以是一个删截样本,即使 y 落入其截断区域,我们也有样本的部分信息;它亦可是一个截断样本,当 y 落入截断区域时则我们不具备任何样本信息。此处我们使用了与赫克曼(Heckman, 1992:205)相同的术语名称,但在文献中,这类术语的使用却并不一致:类似删截随机变量的说法相当常见,其中 c 被称为"删截"(而不是截断)域值。但我认为名称反而是第二位的,读者理解删截数据和截断数据的不同才是重点。

接下来我们将区分两大类删截样本,它们之间的区别在于决定因变量 y 是

否具有确切观察值的机制有所不同。在类似本书列举的第一个例子的一般删截问题里，y 的观察值的特性取决于其本身，例如大于域值 c。但在选择性样本问题中(Heckman，1979)，y_i 是否能被确切地观察，取决于另一变量 z_i 的值。我们可以举一个简单的例子，比如成人给予其孩子零花钱的数额(y)。因为不是所有的成人都有孩子，所以在一个子样本中，我们不具备 y 的观察值。若以 $z_i = 1$ 表示第 i 位成人有孩子，反之 $z_i = 0$，则我们需要两个步骤来解决问题：(a)拟合所有样本中成人有孩子的概率；(b)在有孩子的样本中，拟合 y 的期望值。因而选择性样本是删截问题的一种，但其因变量的截断是因为存在另一变量 z。仍以此为例，我们会有两列解释变量：w，用以解释成人是否有孩子；以及 x，用以解释 y 的观察值。对于所有样本，我们都有 w 和 x 的所有信息，并不管其处于被选择范围之内(同样可以观察到 y)还是之外(没有 y 的观察值)。w 和 x 可能有一些重合变量，甚至有可能完全相同。删截数据与选择性样本数据的区别有时也被称作“外在选择”(explicit)和“内在选择”(incidental)(Goldberger，1981)。

现在我们有了三种类型的样本：删截样本、选择性样本和截断样本。表 1.1 总结了其中的区别。但这三类样本的结构基本相同，而且它们常被共同称作删截问题。用于处理这类问题的统计模型有时也会被总称为 Tobit 模型(Amemiya，1984)，尽管严格而言，Tobit 模型仅是处理这类数据的特殊模型中的一种。

表 1.1　删截样本、选择性样本及截断样本

样本类型	因 变 量	自 变 量
删截样本	y 仅在其值满足某些条件，如 $y > c$ 时，才可获知其确切取值。y 是截断的随机变量	无论 y 是否有确切取值，对于整个样本，自变量 x 都具有观测值
选择性样本	y 仅在另一随机变量 z 满足某些条件，如 $z = 1$ 时，才具有观测值。y 是截断的随机变量	无论 y 是否有观测值，对于整个样本，x 和 w 都可被观测
截断样本	y 仅在其值满足某些条件，如 $y > c$ 时，才具有观测值。y 是截断的随机变量	仅当 y 具备观测值时，自变量才可被观测

两步模型

上述删截数据、选择性样本数据以及截断数据的共同结构决定了它们要使用两步模型。[2]在所有这些情况下，因变量 y 都只在一个子样本(我们称作选择的子样本)中具有观察值。y_i 是否具备完全的观察值(或者说，某一个案是否落入选择的子样本)可以取决于 y_i 本身(删截模型)，也可取决于另一变量 z_i(选择性样本模型)。而这二者与截断模型之间的区别在于：前者含有选择子样本和非选择子样本的信息，而后者则仅含有选择子样本的信息。因而，对于删截数据和选择性样本数据，我们既可拟合选择层(第 i 个个案进入选择的子样本的概率)，也可拟合结果层(选择的子样本中 y_i 的期望值)；而对于截断数据，我们只能拟合后者。

选择两步模型的另一个好处，是它将本章介绍的方法与连续变量的回归分析以及二分变量的分析模型(如 logit 和 probit 模型)联系了起来。选择层本质上是一个二分变量模型(被选择与未被选择)，而结果层则类似于连续变量 y 对一系列解释变量 x 的回归模型。因此，本章结合刘易斯 · 贝克(Lewis-Beck, 1980)与阿肯(Achen, 1982)关于回归模型的论述，以及奥尔德里奇和尼尔森(Aldrich & Nelson, 1984)对离散型变量的分析方法的讨论来构建模型。

另一部分相关文献则是处理时间数据或事件史模型的。在使用这些模型时，我们关注样本成员在移入另一区间之前，在某区间内花费的时间(例如从无业到工作)，以及不同个体在不同时间点发生区间转换的风险。通常这类数据的观察都建立在一个固定的时间区间 T 内。有一些样本成员在这段时间内并未经历转换，因而我们仅仅知道其在原始区间内所花费的时间至少等于 T。这类个案即为删截的。相反，有些样本成员则在时间 T 内经历过转换，则我们会知道其在原始区间内花费的具体时间。这些则是非删截个案。因此，测量在离开某原始区间之前所花费时间的变量则是以 T 而由上截断的。这类似于我们在收入研究中发现的上截断，因而本章中所介绍的方法(需做微小改动)也可用于这类问题。在时间数据的相关文献中，这类方法被称作加速失效模型(Kalb-

fleisch & Prentice, 1980)。它基本上是一种表示原始区间内停留时间的期望长度(或长度的对数形式)的删截回归模型。尽管删截个案与未删截个案之间的区别仍是重点,但现在时间数据的常用方法关注"风险率"(详细介绍见Allison, 1984)。这类模型在统计学中现已成为高度发展的领域,我们在此不再进行讨论,然而我们仍将在第5章讨论删截模型与加速失效模型之间的关系。

社会科学中的删截、选择性样本以及截断问题

为何我们关注删截、选择性样本及截断问题?最直观的原因是它们在社会科学中的普遍性。现代学科对删截数据的最早估计始于托宾的文章(Tobin, 1958),文中介绍了后人所称的 Tobit 模型[3]。使用 735 户的样本数据,托宾分析了持久消费品支出占总可支配收入之比例与两个解释变量之间的关系,包括户主年龄与流动资产占总可支配收入的比例。在他的样本中,有 183 户的因变量取值为 0,因此因变量 y 以域值 $c=0$ 而截断,从而构成删截样本。从此该模型就被社会科学的许多学科使用。例如在政治学中,迪根和怀特(Deegan & White, 1976)用其分析 1973 年休斯敦地方政府官员候选人用于电视广告的开销,其中 40 名候选人中的 24 名开销为零。又如在社会学中,沃顿和拉格(Walton & Ragin, 1990)用其分析债务国公众示威的严重程度。在其 56 个国家样本中,有 30 个国家没有公众示威记录,其因变量取值为 0。

在删截回归方法使用的许多例子中,删截的域值都为 0。这样的例子有家庭资产如公司股份的所有权、酒精消费量、耐用消费品的购买等。但同时也存在域值不为 0 的情况,例如完成全日制教育的年限(其域值为法定最小离校年龄),有最低工资法的国家的收入情况等。然而并不能因为因变量有一个较高或较低的域值(或两个都有),使得至少一部分个案被聚类在一起,就可以认定删截回归模型成为正确的选择。在详细介绍该模型后,我们会继续讨论在什么时候删截回归模型才是适用的。

选择性样本数据的例子大量存在于社会科学研究中。其被广泛运用的领域之一是评估研究,尤其是对劳动力市场项目的影响研究。其中项目参与并不

是一个随机事件，因而参与行为的效果（如收入，或工作机会）研究不仅需要估计进入项目的几率，也需要估计进入项目后的结果。其详细介绍参见巴尔诺，凯恩和戈德堡的著作（Barnow, Cain & Goldberg, 1980）。

学校效应的研究也是选择性样本问题经常出现的领域（Coleman, Hoffer & Kilgore, 1982）。例如，如果我们关注就读于一类学校比之另一类学校的相对效应，则不仅需要研究入选某类学校的过程，还需要分析入选后就读于该类学校的影响。

调查中的无应答也会产生选择性样本问题。如某问卷中关于性生活频率的问题遭遇了大量的无应答。若该无应答是随机的，则使用已应答的子样本来模型化性交频率的解释变量不会存在问题。但很显然无应答并不是这样纯粹随机的，这导致仅对回答者提供的信息进行 OLS 回归可能出现偏误估计。此处我们同样应该进行两步估计：首先是无应答或应答的过程，然后是在应答者中估计其性生活的期望频率。

某些时候选择性样本问题和删截问题可能同时发生。例如对刑事司法系统的研究，若关注对有罪被告人所判处的监禁时间，则我们应关注其中的多阶段过程。首先，在被带入法庭的被告人中，仅有一部分被发现（或辩称）有罪；而在有罪的被告人中，仅有一部分会接受监禁。在第一阶段，我们可以使用选择性样本的方法去估计被判定有罪或宣告无罪的过程，而将第二阶段看作删截样本的例子，因为同样的自变量可以用于决定被告人是否被判监禁以及监禁的时间。因此我们可以对有罪认定的被告人的获刑时间拟合一个删截回归，但应对其做样本选择性偏误修正，因为有些人并不被认定有罪。整个刑事司法程序可以看作不仅是两个步骤的综合，而且是整体一系列阶段的综合（拘留、传讯、审判、判决），其中每一阶段都在上一级样本中选择一个较小的子样本。因而理想状态应是整个过程被拟合为一系列选择性样本和删截数据的模型（Hagan, 1989; Hagan & Parker, 1985; Peterson & Hagan, 1984）。

某些抽样会导致截断问题。如研究者并不总是从总体整体中抽取样本，而是直接在自己所感兴趣的那部分总体中进行抽样，如仅调查那些收入在贫困线以下的家庭。此时如果研究者对收入和教育之间的关系进行测量，则 OLS 回归

必然导致有偏的参数估计。即使研究和所感兴趣的仅仅是贫困家庭中的此项关系，其参数估计仍然有偏误(Berk，1983:388)。两步骤模型(为删截数据所设计)并不能解决这类问题，因为关于贫困线以上的家庭我们没有任何信息，所以只能使用截断数据的分析技术。对找到工作前的无业期的研究同样面临此类问题。因为无业人群并不是总体的随机子集，因而仅根据无业人群的信息所做的参数估计很有可能出现偏误。

理论基础

接下来，我们假设社会科学研究者们希望对一个或多个解释变量(自变量)与某个因变量之间的关系进行总体参数估计，并假设这些估计使用总体的一个随机样本进行。

本书所介绍的删截、选择性样本，以及截断数据的分析方法，如前文所述，包含两个步骤，并对两个步骤分别拟合模型。选择该方法的原因是显而易见的。作为一项标准统计结果，我们可以将随机变量 v 的期望值看作以下两项的乘积之和：v 落入某一分散区间的概率，以及 v 在该区间内的期望值。若以 $I_m(m = 1, 2, \cdots M)$ 表示各分散区间，则 v 的期望值为：

$$E(v) = \sum_{m=1}^{M} \text{pr}(v \in I_m) E(v \mid v \in I_m) \qquad [1.1]$$

其中 $\text{pr}(v \in I_m)$ 表示 v 落在第 m 号区间的概率。因而，随机变量的期望值可以表示为其条件期望 $(E[v \mid v \in I_m])$ 乘以概率 $(\text{pr}[v \in I_m])$ 之和。等式1.1即为“期望的全概率法则”的简单形式(Karlin & Taylor，1975:8)。

将该结论运用于删截问题，因为在一般回归中我们有：

$$E(y_i \mid X_i) = X'_i\beta \qquad [1.2]$$

其中下标 i 表示样本中的第 i 个个案，而 X 和 β 皆为列向量组。

若 y 在一常数 c 两端有两类不同的取值，则根据等式 1.1 的结果，我们可以将等式 1.2 的左侧写作：

$$E(y_i \mid X_i) = \mathrm{pr}(y_i > c \mid X_i)E(y_i \mid y_i > c,\ X_i) + \mathrm{pr}(y_i \leqslant c \mid X_i)E(y_i \mid y_i \leqslant c,\ X_i) \qquad [1.3]$$

此时等式 1.1 中所指的区间由变量 y 自身决定：I_1 是区间$(-\infty,\ c]$而 I_2 是区间$(c,\ +\infty)$。y 是否超过 c 的概率则被看作与 X 相关，而等式的期望值部分则不仅取决于 X，也取决于 y 与 c 的大小关系。因为 y 被 c 分为两部分，所以 y 小于或等于 c 的概率为 1 减去 y 大于 c 的概率。所以等式 1.3 可写作：

$$E(y_i \mid X_i) = \mathrm{pr}(y_i > c \mid X_i)E(y_i \mid y_i > c,\ X_i) + [1 - \mathrm{pr}(y_i > c \mid X_i)]E(y_i \mid y_i \leqslant c,\ X_i) \qquad [1.4]$$

若 y 以 c 为域值由下截断，则其观察值的期望值为：

$$E(y_i \mid X_i) = \mathrm{pr}(y_i > c \mid X_i)E(y_i \mid y_i > c,\ X_i) + [1 - \mathrm{pr}(y_i > c \mid X_i)] \times c \qquad [1.5]$$

注意最后一项条件期望值在等式 1.5 中被常数 c 替代了。这完全无伤大雅，因为若我们定义 $z = y - c$ 并将 z 作为因变量，则我们可以设 $c = 0$。虽然这会改变截距的原始估计值 α，使之变为 $\alpha - c$，但这并不会改变其他的斜率估计值。现在看来，我们仅需估计等式的两个部分，正如之前讨论的那样，估计选择(某一个案不被删截的概率)和结果(未删截个案的条件期望)两个步骤。而这两项都被看作是同一系列变量 X 的函数。

事实上，该模型也并不必要如此严格。选择和结果步骤并不要求是同一列变量的函数。回到等式 1.1，其分散区间也不必以随机变量 v 来定义，我们亦可用另一变量 z 来定义其区间。类似的，选择过程也可能比模型所显示的更复杂。前面我们已经提到了双重截断(同时具有上下限)的情况，同样，我们也可能有更复杂的选择性样本，如仅当两个标准被满足时，y 具有确切的观察值。如我们有两个随机变量 z 和 r，则仅当 $z_i > 0$ 且 $r_i > 0$ 时，我们能观察到 y_i。此时模型为：

$$E(y_i) = \mathrm{pr}(z_i > 0,\ r_i > 0)E(y_i \mid z_i > 0,\ r_i > 0)$$

为了方便表示，我们省去了模型两部分中的解释变量，但 r、y、z 都可以是不同解释变量的函数。如果 r 和 z 互不独立，则选择过程的拟合需要考虑二元概率分布。

更基本的复杂性来自于两步骤的同步而非顺序发生。例如，我们希望研究与人们的收入相关的变量。在工作年龄的成人总体中抽取随机样本，则并非每个人都有工作，因而，因变量收入（或收入的某些转化形式）仅对样本的某些成员具有观察值。若我们进一步假设人们仅仅会从事那些工资在其接受程度的最低限（“保留性工资”）以上的工作，则选择过程（个人是否有工作）和结果过程（当个人有工作时，其工资是多少）并不是顺序发生的。相反，它们同时发生。只有当工作报酬高于个人的保留性工资时，我们才能观察到某人有工作。我们将在第 5 章讨论这类模型。尽管这类同时性使得模型的估计更复杂，但我们仍相信两步骤模型在处理这类问题时的优越性。

本书内容

在下一章我们将首先介绍删截样本的最简估计形式，即 Tobit 模型（Tobin，1958）。我们将用较大篇幅解释与其相关的问题，如最大似然估计和参数解释。在第 3 章，我们会讨论基本的选择性样本数据模型以及截断回归模型。第 4 章通过最大似然估计法丰富删截模型和选择性样本模型，将本书介绍的方法与非连续因变量的其他回归方法如有序 probit（ordered probit）模型相联系。同时，我们会讲述如何将模型扩展为适合选择和结果顺序发生的案例。

第 5 章是关于这些方法所面临的争议及困难，并在模型的现实运用和寻找替代方法方面提供指导。

第 2 章 | 删截数据的 Tobit 模型

处理删截数据的最简单模型是所谓 Tobit 模型(Tobin, 1958),它所处理的即为第 1 章所介绍的问题。我们使用托宾的模型运用的另一个例子来开始讨论。以 y_i 表示第 i 户家庭用于奢侈品的花费,其中第 i 户家庭来自于一个所有家庭的随机样本。以 X_i 表示一系列解释变量的值(如月收入、财富、家庭成员等)。我们需估计向量 β,它包含一系列总体回归参数,表示 X_i 对奢侈品消费的影响。样本包括 N 户家庭,其中 N_0 户家庭不消费任何奢侈品,而另 $N_1(=N-N_0)$ 户则消费某些奢侈品。

删截的潜在变量

Tobit 模型及其他类似模型共同认为存在一个潜在变量 y^* ,y 是其现实观察值。在之前关于考试成绩的例子里,潜在变量为个体学生的真实考试成绩(从 0 到 100),但该潜在变量只有在域值被超过时才可被观测。真实考试分数可用 y^* 表示,而观测值(从 39 截断)则可用 y 表示。类似的,在本章的例子中,y^* 表示家庭在奢侈品上的消费能力或消费倾向,但我们观测到的是家庭的实际消费值 y,它只在消费能力大于 0 时才会出现。所以尽管许多个案的观测值同样为 0,但其潜在变量的取值可能不尽相同。模型的潜在变量形式为:

$$y_i^* = X'_i\beta + u_i \qquad [2.1]$$

假设 u_i 是独立并且服从正态分布的误差项,且其均值为 0,方差为常数 σ^2。重要的是,我们亦假设等式 2.1 是潜在变量与 x 之间的正确关系函数,并且 x 无测

量误差，也不存在任何遗漏变量。所有这些假设都非常重要，因而在使用Tobit模型之前，我们应考虑数据是否满足这些假设。在 OLS 回归中，违反这些假设的后果是非常清楚的，但对本书讨论的 Tobit 模型和其他模型而言，这些后果尚不明确。然而（我们将在第 5 章详细讨论）我们却知道相比于 OLS 模型，这些模型在违反假设（如正态分布假设）时更加不稳健。但这并不说明此类模型过于脆弱而不宜使用：因为很明显 OLS 回归并不适用于这类数据（如示例所言）。坚持使用 OLS 回归，并期望它对于或者真实存在，抑或只是我们怀疑的问题显示出更强的稳健性，这于理不通。更重要的做法是检验这些假设是否被满足，并在可能的情况下转化我们的数据，以使之满足假设，或在研究设计及研究进行过程中最小化这类问题。

观测变量与潜在变量之间的关系可以简单地写作：

$$y_i = y_i^* \quad \text{if} \quad y_i^* > c$$

$$y_i = c \quad \text{if} \quad y_i^* \leqslant c$$

其中 c 为删截的域值（在例子中 $c = 0$）。

若将我们的模型写为观测变量的 y 的形式，并令 $c = 0$，则有：

$$\text{若 } y_i > 0\text{，则 } y_i = X_i'\beta + u_i$$

$$\text{否则 } y_i = 0$$

第 1 章中的等式 1.5 表示某删截于 c 的变量取决于 X_i 的期望值。在本例中，由于 $c = 0$，因而等式可以简化为：

$$E(y_i \mid X_i) = \text{pr}(y_i > 0 \mid X_i)E(y_i \mid y_i > 0,\ X_i) \qquad [2.2]$$

现在我们说明怎样使用两步骤方法去拟合该模型。

两步骤模型

选择

由于 $y_i > 0$，所以：

$$X'_i\beta + u_i > 0$$

因此，

$$u_i > - X'_i\beta \qquad [2.3]$$

换言之，$y_i > 0$ 的概率即为 u_i 超过 $-X'_i\beta$ 的概率。由于 u_i 服从正态分布，则该概率实为某一正态分布的变量超过某值的概率。回忆 z 检验的程序，我们可以从标准正态曲线下的一块面积看出：某一均值为 0，标准差为 1 的正态分布随机变量小于或等于 z 的概率值。现在我们的例子稍有不同，因为我们试图得知 u_i 超过 z_i^* 的概率，其中 $z_i^* = - X'_i\beta$。由于正态分布的对称性，随机变量超过 z 的概率即等于其小于 $-z$ 的概率，即：

$$\mathrm{pr}(u_i > - X'_i\beta) = \mathrm{pr}(u_i \leqslant X'_i\beta)$$

用 $F(X'_i\beta, \sigma^2)$ 表示均值为 0 方差为 σ^2 的正态分布随机变量小于或等于 $X'_i\beta$ 的概率，或简称其为 F_i，则：

$$F_i = F(X'_i\beta, \sigma^2) = \int_{-\infty}^{X'_i\beta} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-t^2/2\sigma^2)\mathrm{d}t$$

该概率等于均值为 0 标准差为 σ 的正态曲线中，从 $-\infty$ 到 $X'_i\beta$ 所占的比例。从而 F_i 等于 $\Phi(X'_i\beta/\sigma)$，或简写为 Φ_i

$$\Phi_i = \Phi(X'_i\beta/\sigma) = \int_{-\infty}^{X'_i\beta/\sigma} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\mathrm{d}t \qquad [2.4]$$

等式 2.4 又被称为标准正态分布函数，它告诉我们标准化后的正态分布随机变量（均值为 0，标准差为 1）小于或等于 $X'_i\beta/\sigma$ 的概率。为标准正态曲线下从 $-\infty$ 到 $X'_i\beta/\sigma$ 所占的比例。

无论是写作 F_i 还是 Φ_i（我们会使用后者），此项概率都可使用 probit 模型来估计（Aldrich & Nelson，1984：48—49）。在 probit 模型中，σ 和 β 并不被单独定义，模型估计的参数为 β/σ，而且为了方便，我们通常假设 $\sigma = 1$（见 Maddala，1983：23）。

结果

若不满足条件 2.3，则对于 y_i 我们会观察到 0 值。因此我们只需在 $y>0$ 的条件下估计 y 的条件期望值。即：

$$E(y_i \mid y_i > 0, X_i) = X'_i\beta + E(u_i \mid u_i > -X'_i\beta) \qquad [2.5a]$$

由于仅当 u 满足条件时，才有 $y>0$，所以在模型中我们不使用 $E(u_i)$（如同在普通 OLS 模型中所做的那样），而是使用条件期望值 $E(u_i \mid u_i > -X'_i\beta)$。因为我们已经假设 u 的非条件期望值为 0，则其条件期望不为 0。因此，事实上我们在回归式中加了额外的一项。现在的问题在于如何估计这一额外项。我们需要知道关于截断的，正态分布的随机变量的期望值。该统计结果在附录 A 中给予说明。由于 u 是正态分布的随机变量，且由 $-X'_i\beta$ 而下截断，则：

$$E(u_i \mid u_i > -X'_i\beta) = \sigma \frac{\phi_i}{\Phi_i} \qquad [2.5b]$$

Φ_i 仍是对 $X'_i\beta/\sigma$ 测量的标准正态函数，而 ϕ_i 则是其对应的标准正态密度函数，即：

$$\phi_i \equiv \phi\left(\frac{X'_i\beta}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \exp \frac{(-X'_i\beta)^2}{2\sigma^2}$$

应注意区分，Φ_i 是概率，而 ϕ_i 则是概率所对应的密度。等式 2.5 中出现的密度与分布函数之间的比值（ϕ_i/Φ_i），被称为逆米尔斯比率，或风险率，常用 λ_i 表示。

$$\begin{aligned} X'_i\beta + E(u_i \mid u_i > -X'_i\beta) &= X'_i\beta + \sigma \frac{\phi_i}{\Phi_i} \\ &= X'_i\beta + \sigma\lambda_i \qquad [2.6] \end{aligned}$$

该等式的估计十分容易。从 probit 模型的选择步骤的结果，我们可以得到观测值 y 大于 0 的估计概率，这即为 Φ。同样，我们可以得到相应的 ϕ 的估计值（通过第 i 个个案[$X'_i\beta/\sigma$]的标准正态密度函数）。对于那些 y_i 大于 0 的个案，我们

用 Φ_i 和 ϕ_i 的估计值来计算 λ_i 的估计值，即逆米尔斯比率。接着就可以使用 OLS 回归，通过拟合非零的 y 值和 X_i 以及估计 λ 值的关系来计算 β 和 σ 的估计值。

$$E(y_i \mid y_i > 0,\ X_i) = X_i'\beta + \sigma\hat{\lambda}_i \qquad [2.7]$$

也可用等式 2.2 来对模型进行估计，它将 y_i 的期望值看作 $y_i > 0$ 的概率与 $y_i > 0$ 时 y_i 之条件期望的乘积。即：

$$E(y_i \mid X_i) = \Phi_i\left[X_i'\beta + \sigma\frac{\phi_i}{\Phi_i}\right] \qquad [2.8]$$

为估计这一等式，我们同样从 probit 模型的结果中得到 Φ_i，即模型的第一部分的值。则 y 的条件期望可由等式 2.5a 得出。通过使用 Φ_i 和 ϕ_i 的估计值（由 probit 模型得来），则我们可以将等式 2.8 简化为：

$$E(y_i \mid X_i) = \hat{\Phi}_i\beta X_i' + \sigma\hat{\phi}_i \qquad [2.9]$$

它同样可以用 OLS 回归来拟合，但这次我们使用的是全部样本数据。

β 的估计值有时被称作赫克曼两步估计量（Amemiya，1984）（Heckman，1976，1979），尽管这一方法相对简单直接，并且易于使用，但同时也面临许多问题。如系数的标准误及 σ 的估计值都是不准确的。我们将使用例子对其作进一步的说明。

最大似然估计

Tobit 模型使用最大似然估计解决这些问题。它不似一般 OLS 方法那样为人熟知，但它的重要性以及在统计和计量学中使用的广泛性（包括删截数据、选择性样本数据和截断数据）使得我们有必要对其基本原理进行解释（见 Aldrich & Nelson，1984：第 3 章；Eliason，1993；Kmenta，1971：174—182）。

首先考虑含有一个自变量（x）和一个因变量（y）的回归模型。该回归的输出结果会向我们提供三个基本参数：截距 α、回归系数 β 以及假设为正态分布且

相互独立的误差项的标准误σ。使用最小二乘法可得到α和β的估计值，同时在满足假设（同方差、残差独立性、残差零和性，以及u和x不相关）的条件下，OLS估计量为最优线性无偏估计量（BLUE）（见 Johnston，1972：第 2 章及第 5 章），意为最小二乘估计量在所有的线性无偏估计量中是最有效（抽样方差最小）的。

还有另外一种估计α、β和σ的方法，即最大似然估计法。其基本原理是：若估计的一系列参数是总体的参数真值，则它们将最有可能产生观测到的样本数据（或严格而言，它们会最经常地产生观测到的样本数据）。对于随机变量y的一系列N个样本观测值y_1、y_2、⋯ y_N，我们提问：若给定一列总体参数值，则其从总体中得出这些具体观测值的可能性多大？而最大似然估计则是尝试所有可能的总体参数值，直至发现某一列参数，其得出具体样本观测值的可能性最大。因而进行最大似然估计的第一步是写出观测到y_i的某一具体模式的似然值。

我们可用较直接的二项分布来解释其做法，并且我们将证明，这即是泛称的 probit 模型的由来。二项分布的随机变量仅有两个可能取值：0 和 1，且其分布可用一个参数来表示，即其均值π。π等于该变量取值为 1 的概率，因而其取值为 0 的概率是$1-\pi$。所以该二项分布随机变量的概率分布为：

$$f(y)=\pi^{y}(1-\pi)^{1-y}$$

它表示该随机变量取某一特殊值（0 或 1）的概率。若我们从此分布中抽取N个值为样本，则其联合概率分布为：

$$f(y_1,\ y_2,\ \cdots\ y_N)$$

它表示我们N个值的样本取某一特殊组合的 0 值和 1 值的概率。若样本个案之间互相独立，则联合概率可看作边缘概率的乘积，即：

$$f(y_1)f(y_2)\cdots f(y_N)$$

用特殊形式来代替f，则：

$$\pi^{y_1}(1-\pi)^{1-y_1}\pi^{y_2}(1-\pi)^{1-y_2}\cdots\pi^{y_N}(1-\pi)^{1-y_N}$$

$$=\prod_{1}^{N}\pi^{y_i}(1-\pi)^{1-y_i} \qquad [2.10]$$

后面的表达式即为似然函数。但为何该看似等同于样本联合概率分布的表达式被称作似然值呢？其原因在于：尽管两者写法相同，但联合概率分布中分布的参数为固定值(π)而 y 为变量，但在似然函数中两者的位置相反：观测值为固定值(y)而分布参数则为变量。一旦我们写出似然函数，则下一步是在给定样本观测值的情况下，找出使该函数最大化的参数值。现实情况中更方便使用的是似然函数的自然对数形式，称为对数似然值，常用 L 表示。对数似然值是似然值的单调变换，因为两个函数将在同一点取得最大值。在本例中，对数似然值为：

$$L = \sum_{i=1}^{N}[y_i \log \pi + (1 - y_i)\log(1 - \pi)] \qquad [2.11]$$

假设我们的样本含有 2000 个观测值，其中 1472 个取值 1 而剩下 528 个取值 0。为了估计未知参数 π(在 0 和 1 之间)，我们将各个可能的值代入等式 2.11。若首先猜测 $\pi=0.5$，则对数似然函数的值为：

$$L = 1472 \times \log(0.5) + 528 \times \log(1 - 0.5) = -1386.29$$

表 2.1 是根据对 π 的不同猜测而计算的不同的 L 值，可以看出 π 在 0.7 时函数值最大，而更精细的研究则会表明对数似然值在 π 为 0.736 时最大。这即为参数 π 的最大似然估计。

表 2.1　π 的不同估计下的对数似然函数值

π 的估计值	对数似然函数值
0.1	−3445.04
0.2	−2468.91
0.3	−1960.57
0.4	−1618.50
0.5	−1386.29
0.6	−1235.74
0.7	−1160.72
0.8	−1178.25
0.9	−1370.86

最大似然法并不仅限于估计参数 π，此时 π 表示样本数据中取值为 1 的观测值的比例。我们可以将 π 看作数据和参数的函数，假设对样本中的每一个 y_i，我们都有相应的 x_i 值，并且 x_i 为连续变量。则 $\pi_i = f(x_i)$，f 是某一函数。在 probit 分析中，我们有：

$$\pi_i = \Phi(\alpha + \gamma x_i)$$

其中 Φ 表示标准正态分布函数，并假设 $\sigma = 1$。

将该表达式代入等式 2.11，则对数似然值为：

$$L = \sum_{i=1}^{N} [y_i \log(\Phi_i) + (1 - y_i)\log(1 - \Phi_i)] \qquad [2.12]$$

其中 $\Phi_i = \Phi(\alpha + \gamma x_i)$。这就是 probit 模型中的对数似然函数（Aldrich & Nelson，1984：51）。最大化该项则可得到参数 α 和 γ 的最大似然估计。当然，此时我们用于寻找 π 的最大似然估计的简单方法已不再适合，最大化对数似然函数需要更为复杂的办法（见 Eliason，1993：第 3 章）。

现在假设 y 不是分类变量或离散变量，而是一个连续变量，我们以此作为第二个例子，仍遵循最大似然估计的基本程序。我们同样寻找使似然函数最大的总体参数值，并在写出似然函数前弄清样本数据的联合概率分布。而其与上一个例子的重要区别在于：对于分类变量或离散变量，我们可用其具体值计算联合概率（换言之，这类随机变量的概率分布函数已有定义），但对于连续变量而言，事实上并非如此。如同对于二分变量，我们可以找出 y 取值为 0 或 y 取值为 1 的概率；但对于连续变量，我们并不能指出 y 取某一特定值时的概率。因此，在似然值中我们不能使用概率分布函数。相反，我们应该用密度函数，宽泛而言，它对连续变量的意义和概率分布函数对离散变量的意义相同[4]。

假设 y_i 的总体围绕其均值呈正态分布，则其密度函数为：

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-[(y_i - \mu)/\sigma]^2}{2}$$

因此似然函数是所有 y_i 的密度的乘积。取其对数形式，则有：

$$L = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(y_i - \mu)^2 \qquad [2.13a]$$

同样的，最大化此函数即可得到 μ 和 σ 的估计值。若我们假设 μ 在不同的样本个体间变化，设 $\mu_i = \alpha + \beta x_i$，且将其代入对数似然函数，则：

$$L = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}[y_i - (\alpha + \beta x_i)]^2 \qquad [2.13b]$$

最大化该表达式则可得到 α、β 和 σ 的最大似然估计值(MLE)。

最大似然估计值有很多很好的属性，但它们只在样本量很大(满足非极端正则条件)时才存在。在统计术语中，最大似然估计值有较好的渐进性质。这与 OLS 统计量略微有一些不同(当然是在满足 OLS 回归条件，如方差、独立误差、零和误差、解释变项与误差项零相关的情况下)。所以，OLS 估计量是无偏的，意即 OLS 估计的参数的期望值等于总体的参数真值，也就是：

$$E(\hat{\theta}) = \theta$$

其中 $\hat{\theta}$ 是 OLS 的参数估计值。而最大似然估计量则并不满足无偏性，只具备一致性。这意味着当样本量增大时，最大似然估计量会越来越趋近于参数的总体真值(Kmenta, 1971:133—134, 181—182)。这并不是说该估计量是渐进无偏估计——意即在极限情况下，当样本量非常大时，参数估计量的期望值等于其真值[5]。最大似然估计量总是具备一致性，但它们也可能是渐进有偏的。但是如同伊莱亚森(Eliason, 1993:20)所指出的，在所有实际应用中，以及在本书所讨论的所有模型中，最大似然估计量实际上都是渐进无偏的。

OLS 估计量的有效性表示使用最小二乘法得到的参数估计的方差总是小于其他线性无偏估计量的方差。而最大似然估计量则是渐进有效的，即只有在样本量很大时，其估计量的有效性特质才得以成立。最后，如果我们假设误差项服从总体正态分布，则 OLS 估计系数也服从同样的正态分布，因而我们可以计算其置信区间，并进行显著性的标准统计检验。最大似然估计则是渐进正态分布的，我们再一次强调：它意味着大样本的最大似然估计服从正态分布，但对于小样本估计而言，事实并不一定如此。最大似然估计量的方差可以很容易地

从逆“信息矩阵”的对角线中获得。信息矩阵是对数似然函数对参数的第二部分偏导矩阵之期望值的负数(参见 Aldrich & Nelson，1984:54；Eliason，1994:20；Kmenta，1971:182)，然而它们是参数的渐进方差，仅在样本量很大时才使用。

最大似然估计是一种普遍而灵活的技术：只要我们能够写出似然函数——它实际上取决于样本数据如何产生的假设——则理论上我们即可估计总体参数。而在现实中我们则需考虑对数似然函数是否有效[6]，其中一个重要的问题是：该函数是否只有一个最大值。若函数有好几个最大值，则参数估计应考虑其起始值。而 Tobit 对数似然方法并不存在这一问题，因为它仅有一个最高点，即广义凹函数(Olsen，1978)。

Tobit 模型的最大似然估计

在写出 Tobit 模型的似然函数之前，我们应考虑样本数据及所做的假设。为使论述具体化，我们以本节开头所谈的奢侈品消费为例。首先假设 μ_i 服从正态分布，不同观测值的误差项彼此独立，而且误差项与解释变量不相关。其次，对样本中的所有家庭，我们知道其是否有奢侈品消费行为。再次，对于其中的 N_1 个未删截个案，我们知道其消费数额。我们使用这三项信息去构建样本整体的似然函数。由于对所有样本，我们皆知道其是否有删截，因此，删截个案对似然值的贡献为：

$$\prod_{0}(1-\Phi_i) \qquad [2.14a]$$

即所有删截个案的被删截概率(等于 1 减去未删截的概率)的乘积。

而未删截个案的贡献为：

$$\prod_{1}\Phi_i \qquad [2.14b]$$

即所有未删截个案的未被删截的概率的乘积。最后，对于未删截个案，我们还知道其具体消费数额，这同样应为似然函数的一部分：

$$\prod_{1}\frac{1}{\sigma}\frac{\phi[(y_i - X_i'\beta)/\sigma]}{\Phi_i} \qquad [2.14c]$$

此为截断正态分布的密度函数。由于 Φ_i 出现在 2.14b 式中的分子部分，亦出现在 2.14c 式中的分母部分（皆作用于未删截数据），于是可互相抵消。将其与式子 2.14a 合并，则似然函数为：

$$l = \prod_{0}[1-\Phi_i]\prod_{1}\phi[(y_i - X_i'\beta)/\sigma]$$

为了估计方便，我们使用其对数形式，即：

$$L = \sum_{0}\log(1-\Phi_i) + \sum_{1}\log\frac{1}{\sqrt{2\pi\sigma^2}} - \sum_{1}\frac{1}{2\sigma^2}(y_i - X_i'\beta)^2 \qquad [2.15]$$

此为 Tobit 模型的完整对数似然函数。注意未删截个案之对数似然的加总部分与 2.13b 式中一般误差回归模型的对数似然函数完全相同。

为举例说明我们所讨论的方法，我们使用模拟数据，其总体为：

$$y_i^* = 1 + 2x_i + \mu_i$$

其中 μ_i 服从均值为 0 标准差为 2 的正态分布。因而总体参数 $\beta = 2$，$\sigma = 2$。从总体中抽取 2000 个个案作为样本，用以估计 β 和 σ。由于此为模拟数据，因而我们可以简单地使用潜在变量 y^* 进行样本回归。则得到估计值 $\hat{\beta} = 2.126$（标准误 $= 0.052$）；$\hat{\alpha}$（截距项）$= 0.927(0.053)$ 以及 $\hat{\sigma} = 2.02$。

若从 0 对样本进行删截，则我们定义 y 为：

$$y_i = y_i^* \quad \text{if} \quad y_i^* > 0$$

$$y_i = 0 \quad \text{if} \quad y_i^* \leqslant 0$$

在我们的样本中，我们给予 y 472 个 0 值。接着我们使用以下四种方法估计 β 和 σ：

A. 使用所有观测进行 OLS 回归

B. 使用非零观测进行 OLS 回归

C. 赫克曼两步骤估计

D. 最大似然 Tobit 估计

表 2.2 是 4 个拟合模型的结果，可以看出它们对 β 和 σ 有着不同的估计值。最显而易见的是 Tobit 模型和使用潜在变量进行回归的 OLS 模型（当然，在实际数据中我们因为无法获知 y^* 而不可能进行此项比较）得出了非常近似的结果。而赫克曼两步法提供的结果也比较相近。相反，使用 y 或者大于 0 的部分 y 进行 OLS 回归所得到的结果则远远偏离于总体 β 和 σ 的值。正如我们在第 1 章所指出的，这些估计是有偏的。对于方法 A，当使用所有的 y 值时，其偏误的来源是显而易见的。等式 2.8 和等式 2.9 给出了 y 的非条件期望的正确模型，而 y 对 x 所做回归得出的系数估计与式 2.9 中得出的部分并不相等，除非对于所有个案都有 $\Phi_i = 1$（因而 $\phi_i = 0$）。但 Φ_i 是某一个案未被删截的概率，由于删截个案总是存在，因而其不可能为 1。所以若等式 2.9 中的系数是无偏的，则方法 A 估计得到的系数是有偏的。

对于方法 B，它仅仅使用正数的 y 值，而等式 2.7 中给出了 $E(y \mid y>0)$。但该式却不能用正数 y 对 x 的估计得到，因为它违反了 OLS 回归的两个中心假设——u 的均值为 0，以及 u 和 x 不相关——因而不能保证估计量的无偏性和一致性。此时，$E(\mu_i \mid \mu_i > -\beta X_i)$ 不等于 0（因为 u 的非条件期望等于 0），相反它会是 X_i 的函数（Maddala，1983：2）。因此，方法 B 的系数是总体参数 β 的有偏估计。正如我们在第 1 章指出的，非删截部分数据的 OLS 回归无法得到无偏参数估计，即使仅仅对总体的未删截部分而言也是如此。

现在关注方法 C 和方法 D。注意在赫克曼 probit 模型中，如前文所述 σ 设为 1。然而 probit 和 Tobit 模型的系数皆可用于计算 Φ_i，因此我们期望其结果相等。由于其概率都计算于 $\beta X_i/\sigma$，因而 Tobit 模型的系数大约应为 probit 模型的 2.022 倍。表 2.2 表明这项关系大致成立，Tobit 模型的系数略小于 probit 模型系数的两倍。

表 2.2 删截数据回归结果(括号中为标准误)

方法	估计量		
	α	β	σ
(A) OLS(所有样本包括 $y_i = 0$)	1.529 (0.043)	1.681 (0.043)	1.676
(B) OLS(仅 $y_i > 0$)	2.085 (0.060)	1.386 (0.054)	1.704
(C) 赫克曼两步法 probit	0.466 (0.010)	1.084 (0.010)	1.000
回归	0.846 (0.266)	2.174 (0.174)	2.178 (0.456)
(D) Tobit	0.929 (0.059)	2.125 (0.056)	2.022 (0.038)
y^* 对 x 的回归:$E(y_i^*) = 0.927 + 2.126x_i$ (0.053) (0.052)			

最后,我们应提早指出赫克曼方法中的第二步 OLS 回归对 σ 及其标准误的估计是错误的。赫克曼方法能给出 β 的一致性估计,但未能给出 σ 的,并且也不能对其提供渐进一致的标准误。因而我们需对其回归结果进行调整。如第 3 章将描述的,这些调整相当直接。而在删截回归模型中,相对于使用最大似然 Tobit 估计,赫克曼两步估计法不具有任何优越性,尤其是当现在后者已经出现在许多电脑程序中的时候。然而,在下文中,两步模型仍将被广泛地用于处理这类问题。

Tobit 模型的参数解释

为解释 Tobit 模型的结果,我们可以从四类期望值的角度入手考察其估计参数。它们是:

(a) 潜在变量的期望值。在 Tobit 模型中,它是:

$$E(y_i^* \mid X_i) = X_i'\beta \qquad [2.16]$$

(b) 超过删截域值 c 的估计概率:

$$\mathrm{pr}(y_i > c) = \Phi\left(\frac{X'_i\beta}{\sigma}\right) \qquad [2.17]$$

(c) 观测值的非条件期望：

$$E(y_i \mid X_i) = \Phi_i \left[X'_i\beta + \sigma \frac{\phi_i}{\Phi_i}\right] + (1-\Phi_i)c \qquad [2.18]$$

(d) 在大于域值 c 的情况下，观测值的条件期望：

$$E(y_i \mid y > c,\ X_i) = X'_i\beta + \sigma \frac{\phi_i}{\Phi_i} + c \qquad [2.19]$$

分清这四项期望值的区别是非常重要的。诸如 LIMDEP 和 SHAZAM 等软件的 Tobit 最大似然估计程序输出的 Tobit 系数，都直接与潜在变量 y^* 相关，因而它们表示 x 变量一个单位的变化对潜在变量的期望值的影响。换言之，对于潜在变量，Tobit 模型的 β 可以用类似于 OLS 模型的 β 来解释。

因此这些系数在用于解释(b)(c)(d)时不能采取同样直接的方式。此时变量 x 的一个单位的变化对因变量的影响不能直接由系数 β 给出，因为一旦 x 改变，则 ϕ 和 Φ 也会改变，而它们皆对 $\mathrm{pr}(y_i > c)$、$E(y)$或 $E(y \mid y > c)$ 等期望值的表达式有影响，因而其偏导数较难计算。从(a)到(d)的偏导为：

$$\frac{\partial E(y^*)}{\partial x_j} = \beta_j \qquad [2.20a]$$

$$\frac{\partial \mathrm{pr}(y > 0)}{\partial x_j} = \phi(z)\frac{\beta_j}{\sigma} \qquad [2.20b]$$

$$\frac{\partial E(y)}{\partial x_j} = \Phi(z)\beta_j \qquad [2.20c]$$

$$\frac{\partial E(y \mid y > 0)}{\partial x_j} = \beta_j\left[1 - z\frac{\phi(z)}{\Phi(z)} - \left(\frac{\phi(z)}{\Phi(z)}\right)^2\right] \qquad [2.20d]$$

此处，z 是 $X'_i\beta/\sigma$ 的值，我们对第 j 个 x 变量求导[7]（为方便省去下标 i），则四项偏导的符号都与 β_j 相同。

如前所述，$E(y^*)$对 x_j 的偏导即为 β_j。它反映了 y^* 与 x 之间的线性关系

(如等式 2.1 所示)。回到本章开始的例子,家庭月收入的系数 β 表示收入的微小变化带来的奢侈品消费倾向的改变程度,对于其他变量,其解释类似。相反,$\mathrm{pr}(y_i > c)$、$E(y)$和 $E(y \mid y > c)$ 对 x_j 的偏导则都取决于 z 的值,因而是非线性的。

对于(b),系数 β 的解释与 probit 模型相似,唯一的不同,是其用 β 除以 σ。这是因为 σ 和 β 在 probit 模型中不能被分别估计,但在 Tobit 模型中其估计是可以分开进行的。因此使用我们的例子,这表示解释变量的微小变化(如收入)对家庭有奢侈品消费的概率的影响。与 probit 模型的结果相似,x_j 的变化对概率的影响在概率为 0.5 左右时最大,在概率接近 0 或 1 时最小(参见 Aldrich & Nelson, 1984:43)。这是因为其偏导等于系数 β 乘以标准正态密度函数 ϕ_i,而 ϕ_i 在相应概率接近于 0 或 1 时趋向于 0,在相应概率等于 0.5 时取其最大值。

$E(y)$的偏导数等于相关系数 β_j 乘以 $\Phi_i(z)$,即某观测值未被删截的概率。该概率越大,则 $E(y_i)$随 x_j 的变化幅度越大。这显然是合理的,因为若 y_i 超过 c 的概率非常小(即大部分 y_i^* 小于 c),则 x_{ij} 的微小变化对 $E(y_i)$的影响将会很小甚至没有影响,因为 y_i^* 仍会小于 c,而 y_i 则仍为 0。这项偏导说明在其他 x 保持不变的前提下,奢侈品消费的观测值随某一 x 变量的变化而发生的改变。

最后,y 大于域值时的条件期望的偏导,表示某一 x 变量的变化对那些有奢侈品消费的家庭的消费值的影响。如公式所示,该偏导等于相关系数 β 乘以括号内的平方项。后者总是为正,且随 z 的增长而增长。因此那些在奢侈品上消费较多的家庭与消费较少的家庭相比,其消费额对变量 x 的变化反应更大。

麦克唐纳德和莫非特(McDonald & Moffit, 1980)对 Tobit 的偏导形式给出了一项有趣且有用的分解。从式 2.8 中 y 的期望值出发,他们注意到其对某个 x 变量的偏导可以写作[8]:

$$\frac{\partial E(y)}{\partial x_j} = \Phi(z)\left[\frac{\partial E(y \mid y > 0)}{\partial x_j}\right] + E(y \mid y > 0)\left(\frac{\partial \Phi(z)}{\partial x_j}\right) \qquad [2.21]$$

即:

$$= \Phi(z) \times \beta_j \left[1 - z\frac{\phi(z)}{\Phi(z)} - \left(\frac{\phi(z)}{\Phi(z)}\right)^2\right] + \left[X_i'\beta + \sigma\frac{\phi(z)}{\Phi(z)}\right] \times \phi(z)\frac{\beta_j}{\sigma}$$

[2.22]

他们指出 Tobit 偏导形式中重要的一点，即 y 的总变化量可以被分解为两部分：一是删截域值之上的观测值的变化，以其处于域值之上的概率加权（等式 2.22 中的第一部分）；二是此项概率的变化，以其观测值的期望加权。这使得 Tobit 效应的分解成为可能。在奢侈品消费的例子中，我们可以将每一个解释变量的效应分解为两项"次效应"：首先是在家庭有奢侈品消费的前提下，解释变量对消费额的影响；其次是解释变量对具有奢侈品消费的概率的影响。使用模拟数据，我们可以使用样本的 x 均值计算麦克唐纳德和莫非特的分解式，得出以下结果：$E(y)$ 对 x 的导数等于 $\Phi_i(x_i = 0.45513)$，乘以 $\beta(= 2.125)$，则等于 1.278。这可以分解为 0.532，即由 Φ_i 计算的 $E(y \mid y > 0)$ 的变化；以及 0.746，即由 $E(y \mid y > 0)$ 的均值计算的超过域值的概率的变化。因而在此例中概率的变化比均值的变化更显著，占 y 的总变化的 58%。感兴趣的读者可以参见麦克唐纳德和莫非特的论文（McDonald & Moffit, 1980），其中列举了该分解方法的几项实际运用。

一个实际例子

霍诺汉和诺兰（Honohan & Nolan, 1993）运用 Tobit 模型研究爱尔兰家庭总财富中金融资产的份额。其样本包含 3089 户家庭，其中 2121 户有金融资产（如股票、债券、储蓄）。他们用家庭总资产（爱尔兰镑）、家庭年收入、家庭区位（城市或农村，以虚拟变量进入模型），家庭户主性别（虚拟变量，若户主为男性则取值为 1），户主是否专业技术雇员，以及户主是否自雇佣人员来解释家庭资产中的金融资产比例。其中，家庭总资产被看作二次项，因而其平方也被加入模型。表 2.3 是对 3089 户家庭分别拟合 OLS 和 Tobit 模型得到的结果。

最令人惊讶的是，OLS 回归中家庭总资产及其平方的系数皆为负数，表明

金融资产份额随家庭总资产的增加而减少。然而,在 Tobit 模型中此效应消失,两项解释变量的系数在统计上都不显著——这项发现相对更合理。相反,在 Tobit 模型中,收入和专业技术雇员的影响更显著。所有家庭中相对较低的金融资产份额(占总资产的 8%),与房屋资产(占总资产的 55%)和农场(25%)的普及形成了鲜明对比。结果显示:仅仅那些可支配收入相对较高的家庭才会选择投资金融资产。

表 2.3 家庭金融资产比例的影响因素(括号中为 t 值)

自 变 量	方 法	
	OLS	Tobit
常数项	6.26(3.3)	−7.47(2.8)
家庭总资产[a]	−0.69(6.0)	−0.18(1.1)
家庭总资产平方	−0.003(2.5)	0.003(1.6)
家庭收入[b]	0.006(2.5)	0.016(5.3)
城 市	6.14(5.9)	7.35(5.1)
男性户主	4.09(3.0)	3.71(2.0)
专业技术雇员	2.92(2.1)	6.20(3.3)
自雇佣者	−4.56(2.3)	−4.87(1.9)

注:a. 单位是 1 万爱尔兰镑。
b. 家庭总的年收入,单位为 100 爱尔兰镑。
资料来源:霍诺汉和诺兰(1993:83)。

表 2.3 中 Tobit 模型的系数应联系潜在变量进行解释,它表示家庭投资金融资产的倾向或能力。因此,收入的系数 0.016 表示家庭收入一个单位的变化对其金融资产投资选择的影响。而该变化对实际的金融资产份额观测值的影响则由等式 2.20c 给出,它取决于其他系数以及家庭在其他变量上的取值。然而,若我们假设 $\Phi(z)$等于有金融资产的观测概率(2121/3089 = 0.687),则收入对金融资产份额的观测值的影响相对较小——等于 0.011。这即为投资金融资产的概率为总体平均值的家庭,在收入发生一个单位变化时金融资产占有比例的期望变化。它仍然大于 OLS 的偏误及不一致估计,即 0.006。若使用 OLS 回归,则霍诺汉和诺兰会低估收入对家庭金融资产投资的影响[9]。

第 3 章 | 选择性样本模型和截断回归模型

Tobit 模型的缺点之一，是它假设同一列变量及参数既决定截断的概率，又决定观测因变量的期望值。本章我们将放松这一假设，使模型两步骤中的变量效应可以不同，且由不同的变量分别决定每一步骤。

克拉格(Cragg, 1971)的模型弱化了 Tobit 模型的这一中心特征。对后者来说，潜在变量超过域值 c 的概率表达式为：

$$\mathrm{pr}(y_i^* > c) = \Phi\left(\frac{X_i'\beta}{\sigma}\right) \qquad [3.1a]$$

而 y^* 关于 X 的期望值为：

$$E(y_i^* \mid X_i) = X_i'\beta \qquad [3.1b]$$

克拉格模型保留了等式 3.1b，但将等式 3.1a 替换为：

$$\mathrm{pr}(y_i^* > c) = \Phi(X_i'\gamma) \qquad [3.1c]$$

比较这两项，则影响概率的变量虽然保持不变，但表达式的两部分(概率和条件期望)有了不同的系数。假设模型的两步骤相互独立，则这两部分系数可以分开估计。克拉格使用该模型分析汽车的购买行为，认为购买汽车与否的决定和购车花费的决定相互独立。芬恩和施密特(Fin & Schmidt, 1984)提供了另一个例子：建筑物发生火灾的概率是楼龄的正函数，但火灾损失则可能是楼龄的负函数。

选择性样本模型

选择性样本模型扩展了克拉格的模型，放松其对模型两步骤互相独立的假

设。模型的基本思路是：结果变量 y 仅当另一变量 z 满足某种条件时才可被观测。所以此类模型的最简单形式将会包含两个步骤：在第一步中二分变量 $z(=0$ 或 $1)$ 决定 y 是否可被观测，仅当 $z=1$ 时，y 具备观测值；而第二步则是在 y 可被观测到的情况下估计其期望值。

以正式形式表示，则令：

$$z_i^* = w_i'\alpha + e_i \qquad [3.2a]$$

$$z_i = 0 \quad 若 \quad z_i^* \leqslant 0$$

$$z_i = 1 \quad 若 \quad z_i^* > 0$$

$$y_i^* = X_i'\beta + u_i \qquad [3.2b]$$

$$y_i = y_i^* \quad 若 \quad z_i = 1$$

$$y_i \text{ 无观测} \quad 若 \quad z_i = 0$$

用语言表达则是：我们观测到一个虚拟变量 z，它是潜在连续变量 z^* 的显现，而潜在变量 z^* 的独立误差项 e 服从正态分布，并且均值为 0，方差为 σ_e^2。当 $z=1$ 时我们可以观测到 y，而 y 是第二个潜在变量 y^* 的显现，其独立正态分布误差项 u 的均值为 0，方差为 σ_u^2。两个误差项的相关系数为 ρ。因此，u 和 e 的联合分布是二元正态的。而两组解释变量 w 和 x 并不必须互不相交，在某些实际运用中，它们可能是同一组变量。若 ρ 被假设为 0，则我们得到克拉格模型。

在实际情况下，误差项之间的相关，通常被认为是由于等式 3.2a 和 3.2b 中共同省略了某一变量。例如我们关注脱离失业状态的人群的收入。模型的选择等式关注脱离失业状态的概率，而结果等式则以已脱离失业状态的人群的收入，或收入的转换函数为因变量。存在同时影响这两步的相关变量，例如“动机”。那些内在动机强烈的人更容易脱离失业状态，同时也更容易获得高收入。但由于其难以测量，所以不被纳入模型。这可能会导致两个误差之间的相关系数不为 0。

但这种理解却是不正确的。相反，我们应将相关性看作模型内生的固有特质。换言之，即使我们不适用样本，不忽略测量 x 和 w[①] 中的某个共同变量，而

① 原文为 x 和 z。——译者注

是使用总体的理论模型，我们依然假设 $\rho \neq 0$。因此任何导致 u 与 e 相关的因素都是内在不可测的。如伯克和雷所言(Berk & Ray, 1982:383):"即使模型被完美拟合，两项误差仍具共变性。两个模型在本质上受到相同的随机干扰(或共变的随机干扰)。"

选择性样本模型被大量用于许多社会科学研究中[10]。若考虑得足够深入，则在任何社会科学数据中都能发现潜在的样本选择过程。如成年人口的随机样本实际上仅仅是出现在抽样框中的成人总体的随机样本，若抽样框为选民手册，则那些未注册选民资格的成人将不被抽样。那么这是否意味着我们应该修正模型估计中的一切偏误呢？一般而言，回答是否定的，除非我们强烈怀疑未注册人口的非随机性。但在19世纪80年代晚期至90年代早期，这种情况确实存在。对选民统一征收人头税的举措，使不注册行为更容易发生在相对贫困的人群中。我们需要对多样本选择过程是否具有重要影响作出判断，而在某些情况下，我们可以对其忽略不计。

等式3.2展示的结构问题在于：使用 $z=1$ 时的观测值简单对 y 进行 x 上的回归，所得到的 β 估计量不仅不一致，而且有偏(我们将在下文证明其原因)。再一次，解决这一问题需采用两个步骤。第一步是估计个案被选择的概率，或者说是对虚拟变量 z 进行变量 W 上的估计。再在个案被选择的前提下，估计变量 y 的期望值。这是在变量 X 上对 y 的估计，并会修正 y 仅在 $z=1$ 时才有观测值的问题。

我们用赫克曼方法拟合该两步模型。使用所有个案，probit 模型估计 $z=1$ 的概率，则得到系数 α 为：

$$\mathrm{pr}(z_i = 1) = \Phi(w_i'\alpha)$$

由于 probit 模型中 α 和 σ_e 不能分开估计，我们假设 $\sigma_e = 1$。

在第二步中我们估计 $z=1$ 时，向量 X_i 决定的 y 的条件期望值。该步骤的推演过程与 Tobit 模型类似(等式2.5a、2.5b及2.6)。

$$E(y_i \mid z=1, X_i) = X_i'\beta + E(u_i \mid z_i = 1) \qquad [3.3a]$$

$$X_i'\beta + E(u_i \mid e_i > w_i'\alpha)$$

为计算等式 3.3a 中的 u 条件期望，我们引用另一统计理论的结果。该结果说明：二元分布中一个变量在以另一变量被删截时，其期望值为：

$$E(u_i \mid e_i > w_i'\alpha) = \rho\sigma_e\sigma_u \frac{\phi(w_i'\alpha)}{\phi(w_i'\alpha)} \qquad [3.3b]$$

该表达式比 Tobit 模型中的更复杂，因为此时我们不再计算 u 本身超过某特定值的条件期望，而是根据另一变量 e 的取值来计算条件期望。将等式 3.3b 代入 3.3a，可得：

$$E(y_i \mid z = 1, X_i) = X_i'\beta + \rho\sigma_e\sigma_u \frac{\phi(w_i'\alpha)}{\phi(w_i'\alpha)} \qquad [3.3c]$$

为了估计这一模型，我们首先使用 probit 模型的结果，对 $z = 1$ 的子样本计算$\frac{\phi_i}{\Phi_i}$（逆米尔斯比率，用 λ_i 表示），然后对于相同的子样本，我们对 y 使用 X_i 和估计的 λ_i 进行 OLS 回归：

$$E(y_i \mid z = 1, X_i) = X_i'\beta + \theta\hat{\lambda}_i \qquad [3.4]$$

以得出 β 和 θ 的估计值。θ 是 ρ 乘以 σ_u 的估计量，由于 $\sigma_e = 1$，所以它等于 u 和 e 的协方差（σ_{ue}）：

$$\theta = \rho\sigma_u = \frac{\sigma_{ue}}{\sigma_e\sigma_u}\sigma_u = \sigma_{ue}$$

讨论这些模型是为了获得 x 对 y 的效应的好的估计。所以如果我们简单地使用观测到的个案，并对 y_i 使用 X_i 进行回归，则等式 3.4 表明：向量 β 的估计一般是有偏的，因为变量 λ 被省略了。因而样本选择性偏误问题在这个意义上等同于模型的错误设定问题，即忽略了某项自变量。然而在两种情况下 OLS 参数 β 也是无偏的：

1. 若 $\rho = 0$，则表示等式 3.4 中的 θ 为 0，因而其可化简为一般的 OLS 回归式。这是选择和结果过程相互独立的情况。

2. 若参数 λ 和某一 x 变量（如 x_k）之间的相关系数为 0，则该变量的

OLS 回归系数 β_k 是无偏的。这是由于遗漏变量对 OLS 回归的影响。若遗漏变量为 λ，则参数 β_k 的偏误等于 x_k 和 λ 之间的相关系数乘以参数 θ。若相关系数为 0，则偏误亦为 0（见 Johnston，1972：168—169；Kmenta，1971：393—394）。

在第 2 章讨论删截回归的两步骤模型时，我们已经说明结果模型中系数的标准误和 σ 的估计都是不正确的。现在的模型同样如此。调整 σ_u 的估计相对容易。定义 $\delta_i = -\lambda_i(z_i + \lambda_i)$，其中 $z_i = w_i'\alpha$。令 s_u 表示赫克曼方法第二步回归中 σ_u 的错误估计，而 S 表示回归的离差平方和：

$$S = \sum_{1} (y_i - \hat{y}_i)^2$$

其求和符号表示对所有 $z = 1$ 的个案求和。而 σ_u 的正确渐进估计为：

$$\hat{\sigma}_u = \frac{1}{N}\left(S - \hat{\theta}^2 \sum_{i=1}^{N} \delta_i\right)^{1/2} \qquad [3.5]$$

其中 N 是 $z = 1$ 的样本个数，θ 是 λ 的估计回归系数（Greene，1990：744—745；Heckman，1979：157）。

标准误之所以错误，则是由于两个原因：模型 3.4 是异方差的，并且使用 λ 的估计值而非 λ 本身，导致系数 β 的标准误需要考虑 λ 的估计误差。而不幸的是误差标准误（OLS）既可大于亦可小于其正确值，因此不能被用作真实标准误的下界。因此，β 和 σ_u 的正确协方差矩阵 V 为：

$$V = \sigma_u^2 (X^{*\prime}X^*)^{-1}[X^{*\prime}(I - \rho^2\Delta)X^* + \rho^2(X^*\Delta W)\sum(W'\Delta X^*)](X^{*\prime}X^*)^{-1} \qquad [3.6]$$

参数估计的标准误由矩阵 V 的对角线的平方根给出。此处 X^* 是矩阵 $[X; \lambda]$；W 是 probit 中解释变量的矩阵；Δ 是对角线为 δ_i、其他为 0 的矩阵；I 为单位矩阵；而 $\sum$ 则是 probit 参数的渐进协方差矩阵。对 ρ 进行估计，有：

$$\hat{\rho} = \frac{\hat{\theta}}{\hat{\sigma}_u} \qquad [3.7]$$

因此,标准误的修正需要一些矩阵操作(Greene, 1981);而一些软件包如LIMDEP(Greene, 1991)则可自动进行这类修正。

模型也可使用最大似然估计,但我们需要定义似然函数。令 $\Phi_i = \Phi(w_i'\alpha)$,则所有 $z=0$ 的个案对似然值的贡献为 $1-\Phi_i$,而 $z=1$ 的个案的贡献为:

$$\Phi_i \times \frac{1}{\sigma}\phi(y_i \mid z_i = 1) \qquad [3.8]$$

其中 σ 是 y^* 在 $z=1$ 时的标准差,而 $\phi(y_i \mid z_i = 1)$ 是 $z=1$ 时 y^* 的条件密度函数。则等式 3.8 实为被选择概率乘以选择样本中 y 的条件密度的表达式。我们需要作进一步的处理以使其更易操作。这超出了本书的范围,但雅美米亚(Amemiya, 1984:31—32)证明等式 3.8 可以写作:

$$\Phi\left[\frac{w_i'\alpha + \rho\left(\frac{y_i - X_i'\beta}{\sigma_u}\right)}{(1-\rho^2)^{1/2}}\right] \times \frac{1}{\sigma_u}\phi\left(\frac{y_i - X_i'\beta}{\sigma_u}\right) \qquad [3.9]$$

加入 $z=0$ 的个案的表达式,并取其对数形式,则对数似然函数为:

$$L = \sum_0 \log(1-\Phi_i) + \sum_1 \log\frac{1}{\sqrt{2\pi\sigma_u^2}} - \sum_1 \frac{1}{2\sigma_u^2}(y_i - X_i'\beta)^2 \qquad [3.10]$$

$$+ \sum_1 \log\Phi\left[\frac{w_i'\alpha + \rho\left(\frac{y_i - X_i'\beta}{\sigma_u}\right)}{(1-\rho^2)^{1/2}}\right]$$

似然函数的作用之一,是告诉我们在何种情况下模型可以得到简化。注意如果 $\rho = 0$,则等式 3.10 可以分为两个部分:一是被选择的概率单位,二是被选择子样本中 y 的 OLS 回归期望值。而由于这两部分并不存在共同参数,所以它们可以分开估计。这表明:若 e 和 u 之间不存在残差相关性,则简单 OLS 回归是合适的。因此,与其说 y 仅对一个选择性样本存在观测值导致估计困难,倒不如说是由于选择的非随机性而导致估计困难。

现在我们有三种可能的方法来分析选择性样本数据:简单 OLS 回归、赫克曼两步估计以及最大似然估计。其中 OLS 回归的估计量既有偏误,也不具有一致性(参见本书第 2 章)。而最大似然估计在满足合适条件的情况下(等式 3.2a

和 3.2b)是渐进无偏和渐进正态分布的，且它会比两步估计更有效。鉴于以上原因，且由于最大似然估计程序的普及，它已成为该模型常用的拟合方法。

例如我们有与等式 3.2a 和 3.2b 所描述的结构相同的数据，在其总体中，

$$y_i^* = 1 + 2x_i + u_i \qquad [3.11a]$$

$$z_i^* = 1 + 2w_i + e_i \qquad [3.11b]$$

且有：

$$z_i = 0 \quad 若 \quad z_i^* \leqslant 0$$

$$z_i = 1 \quad 若 \quad z_i^* > 0$$

$$y_i = y_i^* \quad 若 \quad z_i = 1$$

$$y_i \quad 无观测 \quad 若 \quad z_i = 0$$

其中：

$$\sigma_e: N(0, 1); \sigma_u: N(0, 1.8028); \rho_{e,u} = 0.8321$$

x 和 w 的相关系数为 0.2425。从总体中抽取 2000 个个案作为随机样本，用以估计参数 α(等式 3.11a 的截距项)、β、σ_u 和 ρ。表 3.1 是使用各种方式得出的估计结果。

表 3.1 选择性样本模型估计结果：模拟数据(括号中为标准误)

方法	估计量			
	α	β	σ_u	ρ_{ue}
OLS ($z=1$ 时)	1.2316 (0.0527)	1.9077 (0.0537)	1.7738	
赫克曼两步法 (未修正)	1.0262 (0.0570)	1.9620 (0.0527)	1.7252	0.7266
赫克曼两步法 (修正)	1.0262 (0.0585)	1.9620 (0.0529)	1.7870	0.7014
最大似然法	1.0035 (0.0552)	1.9801 (0.0522)	1.7889 (0.0384)	0.7626 (0.0435)

表格的第一行是使用1173个有具体 y 值的被选择个案进行OLS回归的结果，其中不含有 ρ 的估计，因为此时它被假设为0。第二行和第三行都是赫克曼两步估计的结果，其中第三行提供了修正的标准误及估计的 σ。在这两行中，ρ 都如等式3.7所示计算。最后一行则是最大似然估计。显而易见的，两步估计和最大似然估计皆给出了与总体真值相当接近的 α 和 β，且修正两步估计和最大似然方法都改善了 σ_u 和 ρ 的估计。然而最大似然估计提供了比两步模型更小的标准误。综合四个参数而言，最大似然法提供了最好的估计结果。在第5章，我们将详细讨论这是否为一种必然情况。

参数解释

和Tobit模型类似，该模型的参数估计量也可使用多种方法进行解释。

1. 某个案被选择进入子样本[①]的概率由模型probit部分的系数提供。在最大似然估计中，它们是与其他系数共同估计的。

$$\mathrm{pr}(z_i^* > 0) = \mathrm{pr}(z_i = 1) = \Phi(w_i'\alpha) \qquad [3.12a]$$

此概率对某个 W 变量 w_k 的导数为：(为方便起见省略下标 i)

$$\frac{\partial \mathrm{pr}(z = 1)}{\partial w_k} = \phi(g)\alpha_k \qquad [3.12b]$$

其中 g 表示 $w_i'\alpha$ 的特定值。

2. 潜在变量 y_i^* 的期望值为：

$$E(y_i^* \mid x_i) = X_i'\beta \qquad [3.13a]$$

而其对某一 X 变量 x_k 的导数即为 β_k。注意它并不是对观测值 y 的边际效应的估计，而是对总体期望值的边际效应的估计。

3. 选择性样本中 y 的期望值为：

$$E(y_i \mid z = 1, X_i) = X_i'\beta + \rho\sigma_u\hat{\lambda} \qquad [3.14a]$$

---

① 原文为样本。——译者注

而 y 对 x_k 的导数是：

$$\frac{\partial E(y \mid z=1)}{\partial x_k}=\beta_k-\alpha_k\rho\sigma_u\left[g\,\frac{\phi(g)}{\Phi(g)}-\left(\frac{\phi(g)}{\Phi(g)}\right)^2\right] \qquad [3.14b]$$

该表达式与 Tobit 模型中的等式 2.20d 非常类似，唯一的不同是它包含系数 α_k，表示 x_k 对被选择概率的影响。在 x_k 既属于变量 W 也属于变量 X 时，该部分显然是相关项。但在我们的模拟数据中，若情况并非如此，则其对 y 的影响即为 β_k。若情况如此，则其对 y 的影响可分为两部分：直接影响 β_k，以及间接影响，它是由于 x_k 的变化同时改变着 λ 的估计值而造成的。由于在等式 3.14b 中，$\rho\sigma_u\{g[\phi(g)/\Phi(g)]-[\phi(g)/\Phi(g)]^2\}$ 的值总是为正，因此两项效应的方向不同。所以若问题变量确实对选择概率和结果的期望值皆有正影响，则忽略导数的第二项会夸大其对 y 的影响。

观测变量 y 的非条件期望可以由 Tobit 系数作出解释，但却不适合于选择性样本模型。在 Tobit 模型中，它包括所有固定在删截值上的 y，但在选择性样本数据中，对于未被选择的个案，我们没有任何信息。

一些实际问题

删截数据和选择性样本数据的分析方法并非不存在任何问题，在第 5 章我们会详细论述这一点。另外还有一些实际问题值得我们注意。首先是模型辨识问题。如前文所述，有些赫克曼两步估计在选择和结果机制的估计中均使用同一组解释变量，而另外一些估计则会在结果机制中使用包含所有选择机制解释变量的某组变量。在这种情况下，结果模型的参数仅仅由于 probit 模型的非线性特征才可辨识。若模型是线性的，则由于误差项之间的非零相关系数，模型不会被识别。在两步骤模型中这更为显而易见：如果所有的 w 变量同样出现在 X 中，若选择模型是线性的，则 λ 的估计为部分 x 变量的线性函数。但一般而言，依赖 probit 的非线性去进行模型辨识是不可靠的。更好的办法是对系数做某些限制，如规定选择步骤中的一个变量对结果变量并无影响。尽管我们需

要根据分析中的概念模型来决定哪项限制更合适，但这会保证模型辨识的可能性。在实际的例子中，依赖 probit 的非线性会使我们难以辨识参数估计，从而导致估计的不稳定性。例如在上面的模拟数据中，若以 x 代替等式 3.11 中的 w，则在修正两步骤估计中，参数 α，β 和 θ 之间的样本相关约为 0.8。考察模型参数估计间的相关性是明智的。若模型辨识依赖于 probit 的非线性，则这项考察就显得更为重要。

两步骤模型的另一个问题是关于 ρ 的估计。由于它是两个值的比率，如等式 3.7 所示，则我们不能保证其一定落入 −1 到 1 的区间。我们必须仔细检验模型是否存在任何可能的错误设定问题。

实证例子

最大似然方法在选择性样本模型中的使用远远少于其用于赫克曼两步骤方法。但哈根和帕克(Hagan & Parker，1985)却给我们提供了一个相当好的例子。他们观察对白领工人罪犯判刑的严重程度(用 11 个数字表示不同的严重程度)的影响因素。其中选择性偏误在于：他们的样本包含所有被受理的嫌疑人，而其中仅有 63%被宣告认罪并获刑。在是否被定罪的概率模型中，哈根和帕克的 probit 模型包含 10 个解释变量，其中仅有 3 个显著(Hagan & Parker，1985：309)。在结果等式中，他们使用了完全相同的解释变量。在不修正选择性偏误时，仅有“受理方式”这一变量显著，其对判刑严重程度有很强的正影响(系数为 3.307，标准误为 0.402)。当在结果等式中加入逆米尔斯比率时，许多系数都改变了方向，但仍不显著。变量“受理方式”的系数基本不变(为 3.452，标准误为 0.443)，而逆米尔斯比率本身也不显著，其标准误与系数大小类似(为−2.905，标准误为 2.306)。加入逆米尔斯比率的最大影响是将常数项由 8.63 改变为 5.14。哈根和帕克由此得出结论：“仅观察那些最后定罪的案件并不会使我们的数据分析产生偏误(Hagan & Parker，1985：309)。”

这类在结果等式中包含逆米尔斯比率，但却发现其影响不大，或者逆米尔斯比率的系数(即 u 和 e 的协方差)本身就不显著的例子并不罕见(Allison &

Long，1987；England，Farkas，Kilboune，and Dou，1988；Sanders & Nee，1987）。相反的例子如对男女收入决定因素的研究（Tienda，Smith & Ortiz，1988），其中逆米尔斯比率的估计非常显著。此处选择性偏误的出现是由于研究使用了美国人口普查1970年和1980年样本中工资或薪金收入不为零的子样本。在该研究中，probit等式中包含15个解释变量，其中大部分都是显著的（Tienda等，1988：208）。而结果等式中包含12个解释变量，其中仅有5个与probit等式相同。无论男性或是女性，其逆米尔斯比率变量的参数估计都很显著，因而即使他们未给出修正后的结果等式，该研究仍表明：不对样本选择性偏误进行纠正，将会导致参数估计的偏误。

比较这些研究，我们看到选择性偏误在判刑严重程度的模型中并不成为问题，而在收入模型中则确实需要解决。但这一相反的结论亦可能是因为后者probit模型中自变量的解释力度大于前者，或是因为在Tienda等人的研究中并不依赖于probit的非线性进行模型辨识。我们将在第5章详细讨论这一问题。

截断回归模型

在删截数据和选择性样本数据中，虽然在不满足某种条件的情况下，我们缺乏个案的 y 值信息，但对于所有个案，我们都有关于解释变量的全面信息。因此我们称变量 y 本身是截断的，但其样本却分别是删截的或选择性的。相反，若在不满足条件的个案中，我们不仅缺少关于 y 值的信息，同时也没有解释变量的信息，则样本是截断的。此时两步骤估计不再适合，因为我们缺少可用于分析第一步骤，或称选择步骤的数据。但我们仍然试图拟合结果模型，如下所示：

$$y_i^* = X_i'\beta + u_i \qquad [3.15]$$

其中 u：$N(0, \sigma^2)$。在我们的样本中，仅当 $y_i^* < c$ 时，我们可观测到 $y(=y_i^*)$。若 y 是收入，则 c 可能为收入的贫困线。因而我们需要估计：

$$E(y_i \mid y_i < c,\ X_i) = E(y_i \mid u_i \leqslant c - X_i'\beta) \qquad [3.16]$$

使用附录A中的结果，则：

$$E(y_i \mid y_i < c,\ X_i) = X_i'\beta - \sigma \frac{\phi_i(m)}{\Phi_i(m)}$$

$$= X_i'\beta - \sigma\hat{\lambda}_i(m) \qquad [3.17]$$

$$\text{其中 } m = \frac{c - X_i'\beta}{\sigma}$$

从等式 3.17 可知，若忽略截断，仅对 y 以 x 做回归，则由于未考虑 λ，β 的估计也是有偏的。然而，在前文的两步骤模型中，我们可以从 probit 模型得到逆米尔斯比率的估计值，并将其作为一个新的变量加入结果估计式中；但现在我们却没有信息去做这样的估计。因而两步骤方法不再适合，可行的方法是最大似然估计。确实，该模型的对数似然简单地包含 Tobit 对数似然中关于未删截个案的部分，我们有该部分的信息。未删截的个案对 Tobit 模型中对数似然函数的贡献可分为两部分：未删截的概率（现在我们不能估计），以及等式 2.14c 中截断的正态分布的密度函数。这即为我们会使用的部分。我们用 $\Phi_i(m)$ 代替该式中的 Φ_i，从而得到以下似然值：

$$\prod \frac{1}{\sigma} \frac{\phi[(y_i - X_i'\beta)]}{\Phi_i(m)} \qquad [3.18]$$

由此，可得到对数似然函数：

$$L = \sum \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum \frac{1}{2\sigma^2}(y_i - X_i'\beta)^2 - \sum \log \Phi\left(\frac{c - X_i'\beta}{\sigma}\right) \quad [3.19]$$

其中 c 不一定为常数。我们可以对其添加下角标 i，以表明它在各个个案中可以不同。

截断回归模型较之删截模型和选择性样本模型都更不常见。它通常被用于一些特殊种类的抽样结构。若我们仅对低收入家庭，或者贫困线以下家庭进行抽样，则我们会得到关于家庭收入的截断样本。豪斯曼与怀斯（Hausman & Wise，1997）的著作给出了这类研究的著名例子。

我们再次使用模拟数据，以举例说明该模型。假设父母社会经济地位与子女大学入学考试成绩 y 的关系表达式为：

$$y_i = 75 + 1.5x_i + u_i \quad [3.20]$$

其中 x 为父母社会经济地位，u 服从均值为 0、标准差为 25 的正态分布。我们有 350 名大学生的样本数据，由于大学入学考试的最低录取分数线为 125，因此该数据是左截断的。若忽略这项截断，直接对样本进行简单 OLS 回归，则我们得到 α、β 和 σ 的估计值分别为 102.60(3.36)、1.157(0.05)、22.545(括号中为标准误)。这些估计量既有偏且不一致，并且它们的值皆与(模拟数据中的)总体真值有不小差距。然而，以等式 3.19 中的对数似然函数进行截断回归估计，则得出 α，β 和 σ 的估计值分别为 72.43(6.36)、1.514(0.08)以及 25.721(1.294)，都与其真值非常接近。

截断回归模型中的 β 估计量常用于偏导数解释。它表示当 x 的值发生微小变化时，变量 y 的期望值改变。由于等式 3.20 概括了总体，因而 β 的解释并不只对 y 超过域值的那部分有效，而是适用于整体。

第 4 章 | 基本模型的扩展

删截模型和选择性样本模型相对容易扩展。例如我们考虑结果等式中的因变量为二分变量,而不再是连续变量的情况。假设我们希望研究离婚,由于样本中不是每个个体都曾结婚,因而有些个案并不具备离婚风险,从而构成潜在的选择性偏误。则选择模型处理一个二分变量,其取值分别表示曾结婚与未结婚。而结果模型则关注曾结婚的人群,考察其离婚的概率。另一项可能的模型扩展出现在结果的测量既存在于被选择样本,又存在于未被选择样本的情况。如对劳动力市场项目的评估研究,我们不仅有参与项目的人群的收入信息,也有未参与项目的人群的收入信息。

类似的,我们可以使用更加精细的选择模型。在研究中我们常常会发现,社会过程往往是一系列的连续选择,而参与这一过程的世代在选择序列中人数逐渐减少。教育即是一例:在教育体系的任何一点都有学生退出,因而那些停留至最高阶段,即博士学位的,仅仅只是 4 岁或 5 岁入学时的世代中的很小一部分。同样的道理也适用于刑事司法程序:在所有被逮捕的嫌疑人中,仅有一部分会被受理,而在被认定有罪的人中,也仅有一部分会被判监禁。若结果变量(如被判监禁的时间)出现在这类过程的末端,则应使用一系列连续的样本选择过程对其进行概念化,而最终的步骤或者是某删截变量,或者是选择性样本的结果变量。

因变量的删截性和(或)样本选择性涉及的可能模型范围非常广泛。有时这些扩展模型可以用两步骤方法估计(Amemiya, 1979; Maddala, 1983:第 6 章及第 8 章),而有些则更为复杂。最大似然估计常常是更合适的方法,因为它具备一系列两步骤估计所没有的良好属性。然而,即使我们能够写出任何复杂的

选择性样本及删截模型的对数似然函数，由于现实条件的限制，也很难对其参数进行估计。例如，我们可以写出一个三次或更高次的对数似然函数，但却找不到可用的程序。另外，Tobit 对数似然为严格凹函数，因此只有一个最大值，但对非标准化的似然函数来说，情况并不总是如此。它可能存在局部最大值，从而导致估计最后收敛至非最大似然的危险。在某些特殊情况下，对数似然函数会相对平坦，从而使得收敛过程缓慢，导致参数估计的不稳定性。因而，我们应当谨慎地使用这类方法，从不同起始值进行多次估计，从而避免局部最大值问题；还应小心参数估计剧烈变动或对数似然函数无变化的情况(Eliason, 1994:45)。

本章将集中讨论两类扩展模型。一类是选择过程中因变量有多个域值的删截模型。我们将看到，此类模型证明了删截回归和其他重要计量模型的紧密联系，如研究定序因变量的定序 probit 模型。第二部分将处理一个我们在第一章中已初步接触过的问题，考虑选择过程和结果过程并不顺序发生的情况。它们被看作同时内生于某一特殊过程，是共同发生的。

多重域值的选择过程

在基本删截模型中(Tobit)，仅存在一个域值 c，它对所有个案都为同一常数。然而在第 3 章关于截断回归模型的论述中我们看到：c 也可被看作在不同个案间变化的变量。这只会对对数似然函数带来微小的改变。而模型可以扩展为使用两个到多个域值。例如，我们仅在变量落入上下极限之间时才可观测到其具体值。日用品贸易(Maddala, 1983:160—161)就是这样一个例子，其价格变化的日常范围意味着我们只能观测到潜在变量 y^* 的双重截断部分。又如仅被允许在事先决定的范围(欧洲汇率机制)内变动的汇率，若假设 y^* 为两种货币间的潜在汇率，则我们仅能观测到其落入汇率限制中的部分。

一般而言，我们有：

$$y_i^* = X'_i\beta + u_i \qquad [4.1]$$

$$u_i : N(0, \sigma^2)$$

并且：

$$y_i = y_i^* \text{ 若 } c_1 \leqslant y_i^* \leqslant c_2$$

$$y_i = c_1 \quad \text{若} \quad c_1 > y_i^*$$

$$y_i = c_2 \quad \text{若} \quad c_2 < y_i^* \qquad [4.2]$$

此时似然函数有三个部分。一是潜在变量在低域值 c_1 之下的个案，其贡献为变量不超过该域值的概率；二是超出高域值 c_2 的个案，其贡献为变量超过该上限的概率；其三是我们有 y 的具体取值的个案，其贡献等于落入域值之间的概率乘以 y^* 的条件密度函数。再次使用附录 A 的结论，则潜在变量 y^* 超过域值 c_m 的概率为：

$$\begin{aligned}\text{pr}(y_i^* > c_m) &= \text{pr}(X_i'\beta + u_i > c_m)\\ &= \text{pr}(u_i > c_m - X_i'\beta)\\ &= 1 - \Phi\left(\frac{c_m - X_i'\beta}{\sigma}\right)\end{aligned}$$

令：

$$\Phi\left(\frac{c_m - X_i'\beta}{\sigma}\right) = \Phi_i(c_m)$$

以简化等式。在本章中，我们将一直使用该缩略形式。

则 $y_i^* > c_1$ 的概率为 $1-\Phi(c_1)$，$y_i^* < c_2$ 的概率为 $\Phi(c_2)$。这两项表达式都将出现在对数似然函数中。仅当 y^* 在两个域值之间时，我们可以确切观测 y^*。则其概率为 $y_i^* < c_2$ 的概率减去 $y_i^* < c_1$ 的概率，即：

$$\text{pr}(c_1 \leqslant y_i^* \leqslant c_2) = \Phi_i(c_2) - \Phi_i(c_1)$$

对这些个案，我们还需要知道 y^* 的条件密度函数，它的分母是同上的表达式。因而在化简后，那些 y^* 可被精确观测的个案，其贡献与在简单 Tobit 模型中一样。因此，完整的对数似然函数为：

$$L = \sum_{y_i = c_1} \log[\Phi_i(c_1)] + \sum_{y_i = c_2} \log[1 - \Phi_i(c_2)]$$

$$+\sum_{y_i=y_i^*}\log\frac{1}{\sqrt{2\pi\sigma^2}}-\sum_{y_i=y_i^*}\log\frac{1}{2\sigma^2}(y_i-X_i'\beta)^2 \qquad [4.3]$$

该对数似然函数与普通 Tobit 模型十分类似。我们可以根据模型计算 4 项有用的期望值。首先是 y 在两个域值之间的条件期望：

$$E(y_i \mid c_1 \leqslant y_i \leqslant c_2) = X_i'\beta + E(u_i \mid c_1 - X_i'\beta \leqslant u_i \leqslant c_2 - X_i'\beta)$$

$$= X_i'\beta + \sigma_u\left[\frac{\phi_i(c_1)-\phi_i(c_2)}{\Phi_i(c_2)-\Phi_i(c_1)}\right] \qquad [4.4a]$$

在前文汇率的例子中，该项表示在变动幅度之内汇率的条件期望值。该表达式使用双重截断的随机正态分布变量的标准统计结果。为得出 u 的条件期望，我们回忆附录 A 中单边截断变量的条件期望为：

$$E(u_i \mid u_i \leqslant c - X_i'\beta) = \sigma\frac{-\phi_i(c)}{\Phi_i(c)}$$

由于 $c_2 - X_i'\beta$ 大于 $c_1 - X_i'\beta$，于是：

$$\mathrm{pr}(c_1 - X_i'\beta \leqslant u_i \leqslant c_2 - X_i'\beta) = \mathrm{pr}(u_i \leqslant c_2 - X_i'\beta) - \mathrm{pr}(u_i \leqslant c_1 - X_i'\beta)$$

$$= \Phi_i(c_2) - \Phi_i(c_1) \qquad [4.4b]$$

这即为双重截断变量 u 的条件期望的分母，则分子为：

$$\sigma\{-\phi_i(c_2) - [-\phi_i(c_1)]\}$$

将分子分母组合，并重新整理，则得到等式 4.4a 中的后半部分。

扩展等式 1.1，可得到 y 的非条件期望：

$$E(y_i) = \mathrm{pr}(y_i = c_1)c_1 + \mathrm{pr}(y_i = c_2)c_2$$

$$+ \mathrm{pr}(c_1 \leqslant y_i \leqslant c_2) \times E(y_i \mid c_1 \leqslant y_i \leqslant c_2)$$

用语言表达则是：观测值 y 的期望值等于其三类个案（y 等于两个极限，或 y 落入这两者之间）的条件期望以概率加权后的总和。在例子中它即为观测汇率的简单期望。

由于该表达式的最后一部分与等式 4.4a 相同，则化简后可得：

$$E(y_i) = \Phi_i(c_1) \times c_1 + [1 - \Phi_i(c_2)] \times c_2 + [\Phi_i(c_2) - \Phi_i(c_1)] \times X_i'\beta + \sigma_u[\phi_i(c_1) - \phi_i(c_2)] \quad [4.4c]$$

第三是潜在变量 y^* 的期望值，即 $X_i'\beta$，所以关于潜在汇率，β 可以使用普通的偏导数来解释。最后是我们可以计算三项期望概率：超过低域值的期望概率为 $1-\Phi(c_1)$；不超过高域值的期望概率为 $\Phi(c_2)$；而等式 4.4b 则是落入两个域值之间的期望概率。参数 β 对这些概率的解释与普通 Tobit 模型相同(第 2 章)。

尽管该模型有其本来的用处，但本书介绍它的主要原因却是展示其与社会科学研究者感兴趣的其他模型的紧密联系。当搜集连续变量数据(如收入)且将最高收入和最低收入综合时，我们可能用到上文描述的模型。然而，典型的状况是，在调查中我们并不会问被访者的具体收入，而是给其一系列收入范围并询问其所处的区间。只要对等式 4.3 做微小的改动，我们就可以用这类数据估计收入(见 Stewart, 1983)。此时我们不具备任何关于 y^* 的信息，我们仅知道被访者的收入大于某值且小于另一值。即，

$$y_i = 0 \quad 若 \quad y_i^* < c_1$$

$$y_i = 1 \quad 若 \quad c_1 \leqslant y_i^* \leqslant c_2$$

$$\cdots$$

$$y_i = M \quad 若 \quad y_i^* < c_M$$

其对数似然函数由观测到 y 的每个取值的概率构成。若我们假设潜在变量 y^* 与向量 X 间的关系符合式 4.1[11]，则其对数似然函数为：

$$L = \sum_{y=1} \log[\Phi_i(c_1)] + \sum_{y=2} \log[\Phi_i(c_2) - \Phi_i(c_1)] + \cdots \sum_{y=M} \log[1 - \Phi_i(c_M)] \quad [4.5]$$

向量 β 表示 x 变量与潜在连续收入变量 y^* 之间的关系。

若假设我们仅知样本中收入的等级序列，则模型可以进一步被扩展。通过改良等式 4.5，可以导出"定序 probit 模型"(McKelvey & Zavoina, 1976; Maddala, 1983:46—49)。其中的域值，或称截点，是我们估计的参数。由于我们不再有足够

的信息去单独估计 σ,所以定义 $d_m = c_m/\sigma,\ m = 1,\ \mathrm{K}\ M$,以及 $\lambda = \beta/\sigma$,则:

$$\Phi(d_m) = \Phi(d_m - X'_i\gamma)$$

若以该表达式替换等式 4.5 中的 $\Phi_i(c_m)$,则为定序 probit 模型的对数似然函数。此处需要估计的参数为 γ 和 d_m。若 X 中包含一个常量,则有一个 d_m 的值不会得到估计。例如,我们有 4 个截点分成的 5 个区间,则 X 中的常量使我们只需估计其中的 3 个。而第一区间会是从 $-\infty$ 到 0,第二区间为 0 到 d_1,直至第 5 区间 d_3 到 ∞。等式 4.5 中的模型显然可被广泛运用于连续变量被准确测量或以区间搜集信息的情况,因此定序 probit 模型在因变量是定序的,且其值实为某一潜在正态分布变量的显现时可被广泛运用。

内生性选择和结果

试想我们有一个成人样本,其中有些人有工作而有些人没有工作。我们希望建立工资与变量 x 之间关系的模型。样本总体符合:

$$\log(wage) = y_i^* = X'_i\beta + u_i \qquad [4.6]$$

我们对 u 做常规性假设——即服从均值为 0 方差为 σ_u^2 的正态分布,并假设样本个案相互独立。仅当个体有工作时,我们观测到 y_i^*。假设所有个体都有其保留性工资 v_i^*,仅当其收入高于或等于 v_i^* 时,他们才会接受工作。v^* 无法直接观测,但可以被看作是:

$$v_i^* = W'_i\alpha + e_i \qquad [4.7]$$

我们对误差项 e 做常规假设,而 W 则是一组可观测的变量。假设 $\rho(e,\ u) \neq 0$。

若个体有工作,则定义 $z = 1$,否则定义 $z = 0$,以及:

$$y_i = y_i^* \quad 若 \quad z_i = 1$$

$$y_i\ 无观测 \quad 若 \quad z_i = 0$$

当使用两步骤方法进行估计时,我们会遇到一些困难。由于:

$$\mathrm{pr}(z_i = 1) = \mathrm{pr}(y_i^* > v_i^*) = \mathrm{pr}(y_i^* - W'_i\alpha > e_i) \qquad [4.8]$$

则是否接受工作的决定取决于潜在工资 y^* [12]，因而我们不能将接受工作看作发生在获得工资之前的选择过程。相反，这两个步骤是同时发生的。

模型的似然函数包括两个部分。那些不工作的个案，其贡献为概率 $v^* > y^*$，即：

$$\begin{aligned}\mathrm{pr}(v_i^* > y_i^*) &= \mathrm{pr}(W'_i\alpha + e_i > X'_i\beta + u_i) \\ &= \mathrm{pr}(e_i - e_i > X'_i\beta - W'_i\alpha) \qquad [4.9]\end{aligned}$$

由于 $e-u$ 服从正态分布，且其方差为：

$$\sigma^2 = \sigma_u^2 + \sigma_e^2 - 2\sigma_{ue}$$

则等式 4.9 可写作：

$$\Phi\left(\frac{X'_i\beta - W'_i\alpha}{\sigma}\right) \qquad [4.10]$$

而似然函数的另一个部分则是关于那些有工作的个案。或者说满足等式 4.8 中的要求的个案。在之前讨论选择性样本的似然函数时，我们注意到可观测 y 的贡献为其被选择的概率乘以条件密度函数。此处的情况稍微复杂一些。我们需要看 u 和 e 的二元密度函数，且 e 在 $y_i^* - W'_i\alpha$ 处截断，即：

$$\int_{-\infty}^{y_i^* - W'_i\alpha} f(u_i,\ e_i)\mathrm{d}e \qquad [4.11]$$

其中 $f(a,\ b)$ 表示两个正态随机分布变量 a 和 b 的二元密度函数。等式 4.11 表明，我们现在关注的是当 $e < y_i^* - W'_i\alpha$ 时，u 和 e 的联合密度函数。而该条件则恰好是等式 4.8 中说明的，将有观测工资与没有观测工资的人群分开之条件。

马黛娜(Maddala，1983：76)将这部分似然函数化简得到：

$$\sum_{z=1}\left[\log\frac{1}{\sqrt{2\pi\sigma_u^2}} - \frac{1}{2\sigma_u^2}(y_i - X'_i\beta)^2 + \log\Phi\left(\frac{\sigma_u^2}{\sigma_e^2\sigma_u^2 - \sigma_{eu}^2}\left(y_i - W_i\alpha - \frac{\sigma_{eu}^2}{\sigma_u^2}(y_i - X'_i\beta)\right)\right)\right] \qquad [4.12]$$

其中 σ_{eu}^2 是 u 和 e 协方差的平方。

完整的对数似然函数则是等式 4.12(所有有工作的个案)和等式 4.8(所有没有工作的个案)的加总。为了辨识模型,u 和 e 的相关系数必须为 0,且必须存在一个变量,它在 X 中但不在 W 中。

尽管顺序两步模型在此处不再适用,但似然函数仍包含未选择个案和被选择个案的部分。而其对数似然函数则与等式 3.10 中的基本选择性样本模型类似。其复杂性来源于我们处理了两个内生性的变量,其中一个为截断变量。自格朗纳(Gronau, 1974),刘易斯(Lewis, 1974)和赫克曼(Heckman, 1974)之后,使用这类对数似然函数的模型被扩展应用于劳动力供给问题。正如马黛娜(Maddala, 1983:200—202)所言,它也可以被很容易地扩展至截断和样本选择性问题。

第 5 章 | 应注意的问题

本书的前几章介绍了删截、选择性样本和截断数据模型。毫无疑问，自20世纪70年代末期以来，这些模型在社会科学研究中就开始被广泛地使用。然而近年来却有证据表明这些方法本身亦存在问题，尤其是赫克曼模型，受到了大量的批评。尽管许多方面仍无定论，但显然，我们在使用这类模型时应更加谨慎——如标题所言，本章讨论删截模型和选择性样本模型在实际应用中的三个重要问题。首先是其对分布假设的敏感性；其次是赫克曼模型的辨识问题；最后我们将讨论评估研究中选择性样本模型的运用。我们将不仅说明方法存在的问题，而且提供解决问题的建议或可使用的替代性方法。最后本章将以一系列指导作为小结，我们希望其可避免这些可能的缺陷。

对分布假设的敏感性

异方差

异方差，或者误差项的非常数方差问题，在删截模型和选择性样本模型中比在 OLS 回归中更重要。这是因为在异方差的情况下，最小二乘估计量虽然不是有效的，但却具有一致性。而删截模型和选择性样本模型的估计量却既不一致也不有效(Amemiya, 1984:23)。其解决办法是“对异方差的本质做一些合理假设”(Maddala, 1983:179)[13]。换言之，我们根据异方差的函数形式，将对数似然中的 σ 写作可观测的变量的函数。伊莱亚森展示了在异方差时误差项的正态分布假设(Eliason, 1993:28—34)和截断正态分布假设(Eliason, 1993:63—66)下可做的调整。马黛娜(Maddala, 1983:180)则建议对 Tobit 模型中的误差

项假设：

$$\sigma_i = (\gamma + \delta Z_i)^2$$

此时 Z 为向量 X 中的部分或所有变量，而 γ 和 δ 则是需要估计的参数。我们用其替换式子 2.15 和 Φ_i 表达式中的 σ。

非正态性

样本选择偏误方法本身(与其具体执行过程相区别)并不对分布进行严格假设(见 Heckman & Robb，1986：57—63)。例如赫克曼原始两步估计量只要求(a)选择式的误差服从正态分布；以及(b)结果变量条件期望等式的误差是选择等式误差的线性形式(Olsen，1980：1817)。然而这些标准方法的执行过程(两步骤方法或最大似然估计法)却需要对结果等式误差以及两误差联合分布(如二元正态分布)进行假设。

删截模型和样本选择模型中的非正态性具有极大的潜在危害。尽管 OLS 估计量在非正态条件下也是一致的，但选择性样本模型和删截模型的估计量则不是。戈德堡(Goldberger，1983：79)测量了 Tobit 模型中一系列对称但非正态的误差分布，得出的结论是："一般样本选择性偏误的修正模型通常会对偏离正态性十分敏感。"更一般的，选择性样本偏误修正方法的研究者奥尔森(Olsen，1982)指出："最大似然估计方法由于对回归残差的总体分布假设过于敏感，因而不具备优良特质。"

既然非正态性是潜在的如此严重的问题，则我们应如何对其进行处理呢？主要有两种办法：若我们有依据为误差项假设一个已知的参数分布，则我们可将该假设分布纳入模型；而若误差分布是完全不可知的，则我们应使用半参数方法。下面我们分别介绍两种方法。

关于非正态误差的最直接的参数方法，是最大似然估计法，其中我们直接定义误差项的分布。马黛娜(Maddala，1983：187—190)列举了非正态 Tobit 模型的两个简单例子。其一是假设误差项服从对数正态分布。此时模型的对数似然函数与一般模型的大致相同，我们只需用 $\ln(y)$ 替换 y，并使用域值的对数

形式即可。其二是假设 u 为指数分布，其密度函数和分布函数皆为非常简单的形式。Tobit 模型由于只涉及单变量分布，因而相对较易扩展到非正态分布的形式。正如格林所指出的（Greene，1991：588），生存分析中使用的加速失效时间模型实为删截回归模型，但其通常具有非正态的误差分布。典型的分布有韦氏分布（Weibull）、对数逻辑斯蒂分布（log-logistic）、Gompertz 分布等（Allison，1984），因而软件可以较容易地估计非正态 Tobit 模型。

对于最大似然选择性样本模型来说，情况则更为复杂，因为我们需处理选择式和结果式两项误差的二元分布。李（Lee，1993）给出了一个操作性的例子。原始赫克曼 Probit-OLS 方法要求选择等式误差的正态性。然而在奥尔森（Olsen，1980）早期工作的基础上，李（Lee，1983）却说明更为灵活的两步骤方法可以估计非正态误差模型（尽管我们需要对其分布做出具体假设）。他的模型相当简单：先在假设误差项分布的基础上计算选择模型，然后计算预测概率，接着找出这些预测概率的逆正态分布函数（也就是计算代入 $\Phi(\cdot)$ 能得到预测概率的值 $W_i'\alpha = j_i$），最后用 j_i 计算正态密度和分布函数，以估计 λ_i。

这在选择等式的结果多于两类时——即在所谓的多项选择模型中，也非常有用。例如，假设我们关注四类学校中学生的数学成绩。我们拟合一个选择等式来解决这一问题，其中包含四个学校类别 ($m = 1$ K 4)，而在结果等式中，学生的数学成绩作为因变量是在四类学校中被分别观测的。李的方法是使用多元 logit 回归来估计学校的选择问题，即计算学生进入每一学校的概率，再用其计算 j_{im}，$m = 1$ K 4。这些数值会被用于估计相应的 λ_{im}，然后被用于四项 OLS 回归[14]。

事实上，李（Lee，1983）的方法比我们的介绍更具一般性。它不仅允许选择等式误差 e，同样也允许结果等式误差 u 为非正态分布（但需为已知分布）。此时逆正态分布函数能将 e 和 u 转化为正态的，并使其联合分布为二元正态的。李（Lee，1983）和马黛娜（Maddala，1983：272—275）对此进行了详细的论述。这使我们可以使用最大似然估计法。由一般选择等式和结果等式出发：

$$z_i^* = W_i'\alpha + e_i \tag{5.1a}$$

$$若\ z_i^* \leqslant 0, 则\ z_i = 0$$

$$若\ z_i^* > 0, 则\ z_i = 1$$

$$y_i^* = X_i'\beta + u_i \qquad [5.1b]$$

$$若\ z_i = 1, 则\ y_i = y_i^*$$

$$若\ z_i = 0, 则\ y_i\ 无观测$$

假设 e 和 u 的相关系数为 ρ。现在假设 u 的密度函数为 $g(u)$ 而累积分布函数为 $G(u)$，则 e 的累积分布函数为 $F(e)$。F 和 G 不服从正态分布。若设标准正态分布函数的逆函数为 Φ^{-1}，则我们引入新变量 e^*，使：

$$e^* = \Phi^{-1}[F(e)]$$

这即为我们上文描述的转化过程：找到一个值，使其代入标准正态分布函数时可得到概率值 $F(e)$。对 u 进行同样的转化，有：

$$u^* = \Phi^{-1}[G(u)]$$

e^* 和 u^* 皆为标准正态分布。而其二元分布

$$B\{\Phi^{-1}[F(e)],\ \Phi^{-1}[G(u)],\ \rho\}$$

由于 e^* 和 u^* 的正态性，会服从二元正态分布。我们简单的将上述表达式代入式 3.10 的选择性样本模型，则用最大似然法估计等式 5.1a 和 5.1b，有：

$$L = \sum_0 \log[1 - F(W_i'\alpha)] + \sum_1 \log\{g[(y - X_i'\beta)/\sigma_u]\} \qquad [5.2]$$
$$+ \sum \log \Phi\left[\frac{\Phi^{-1}[F(W_i'\alpha)] + \rho\{\Phi^{-1}[G(y - X_i'\beta)/\sigma_u]\}}{(1-\rho^2)^{1/2}}\right]$$

此处 $F(W_i'\alpha)$ 是 e 小于 $W_i'\alpha$ 的概率，而 $G(y - X_i'\beta)/\sigma_u$ 则是 u 小于 $(y - X_i'\beta)/\sigma_u$ 的概率。最后一部分中的 Φ^{-1} 计算当代入 $\Phi(\cdot)$ 可以得到相同概率的标准正态分布的值。若 e 和 u 为正态分布，则可化简为式 3.10，因为 g 是正态密度函数，F 和 Φ 相同，最后等式 5.2 中的 F 和 G 都能与 Φ^{-1} 抵消。实际上，该方法可以非常灵活的转化非正态分布的误差项，使其可用于删截模型和选择性样本模型的

一般方法(尽管在一些情况下,我们需要引入一些附加的限制条件,使得模型估计的误差处于我们所选择的特殊分布允许的范围内)。

若误差项的分布是未知的,则我们应使用半参数模型。在赫克曼方法中,我们曾估计向量 α,并用其找出这样一个表达式:

$$E(\text{结果等式误差}\mid\text{选择等式误差} > W_i'\alpha)$$

纽维尔、鲍威尔和沃克(Newey, Powell, & Walker, 1990)说明了两种估计 α 的半参数方法,以及两种以 α 的估计值为基础,在结果等式中估计 β 的半参数方法。然而如马黛娜(Maddala, 1992:56)所言,这些非参数方法还比较初步,其实际运用仍较罕见。所以本书不再赘述这些估计量的推导过程,有兴趣的读者请参考纽威等人的著作(Newey 等,1990)。科斯莱特(Coslett, 1991) 也给出了选择性样本模型的一种非参数估计;鲍威尔(Powell, 1984)则提供了非参数估计的 Tobit 模型。而李则论述了删截模型和选择性样本模型的一般非参数估计[15]。

鉴于这些模型的一般估计量对偏离正态性和同方差的敏感性,因而对这些假设的检验显然是非常有必要的。学者们提供了一些这方面的方法(如 Lee and Maddala, 1985)。有一类对正态性和同方差的偏离检验是针对特定假设进行的。切希尔和艾利时(Chesher & Irish, 1987)提供了一组非常有用的检验方法。而这些检验方法并不需要设定异方差和非正态分布的具体形式。例如它们只是根据偏度和峰度来建立标准化的正态性检验,在删截模型中,由于潜在变量只能被部分地观测,因而对标准化残差项的检验也不是以直接观测为基础的。

对其方法感兴趣的读者可以参见其论文。其基本思路是检验潜在变量 y^* 的标准化(均值为 0、标准差为 1)回归残差的估计分布矩,并与正态分布假设下其应有的值,也就是标准正态分布的矩相比较。该检验主要有三步:首先是在 y^* 仅被部分观测的基础上计算标准化残差;然后是计算这些残差与正态分布残差的矩的差值(这些值被称为矩残差,最多会有四项矩:均值、方差、偏度和峰度);最后进行一项评分检验,以考察该差值的显著性,或者更准确地说,是检验

残差的观测分布与正态性假设下的分布之无差别的零假设。同方差检验的方法也遵循类似的原理。

这些假设可以被简化为一个简单回归，因而很容易计算。在同方差检验中，前两项"矩残差"是必要的，而在正态性检验中，我们则需要计算总的四项。它只涉及观测值 y、解释变量以及 β 和 σ 的估计量，所以并不复杂（在附录 B 中我们介绍了计算方法）。这些矩残差和解释变量一起构成新矩阵 R。我们用该矩阵对向量一进行回归，则被解释的平方和可用于检验零假设（同方差或正态性）。我们只需做一个简单的卡方检验即可。附录 B 详细解释了该方法。博拉、嘉可和李（Bera，Jarque & Lee，1984），以及戴维森和麦克卡农（Davidson & Mackinnon）也提供了正态性检验的类似方法。

模型辨识和稳健性

许多学者（如 Little，1985）都意识到在原始 probit-OLS 模型中依赖非线性来完成模型辨识所可能带来的问题。如伯克和雷指出的一系列由此产生的问题：

> 典型的结果是高方差的估计量。而真实（结果）等式中风险率指标与其他回归量之间的多重共线性也是一个常见的问题……最后，若无法解释选择过程中的大部分方差……则风险率（逆米尔斯比率）的方差将会很小……（这将导致）真实等式中与截距项的高度相关。（Berk & Ray，1982:386）

杜安、曼宁、莫里斯和纽豪斯（Duan，Manning，Morris & Newhouse，1984:288）给出了多重共线性的一个例子。他们发现逆米尔斯比率和结果等式中的其他解释变量间的多元相关平方 R^2 在样本总数 9 中超过了 0.8。而在对哈根和帕克（Hagan & Parker）的讨论中，我们注意到逆米尔斯比率系数的标准误已经接近于其系数本身，因而其对结论最明确的影响就是将常数项降低至错误值的约 2/3。

斯托增伯格和雷利(Stolzenberg & Relles, 1990)的蒙特卡罗研究，发现即使在选择式误差和结果式误差的二元正态分布成立时，赫克曼两步骤方法也存在严重的问题。使用严重删截的(90%)500个模拟个案，他们发现：赫克曼方法在相关参数估计的偏误和准确性上与OLS回归一般无异。他们由此推断，赫克曼方法在测量和修正样本选择性偏误中作用微小，不宜被经常使用。

斯托增伯格和雷利的文章对使用赫克曼方法修正样本选择性偏误的倾向敲响了警钟(见Land & McCall, 1993)。然而其结论与尼尔森(Nelson, 1984)的早期蒙特卡罗研究大不相同。后者认为赫克曼两步骤技术的问题可以很容易澄清。

尼尔森的文章比较OLS回归，赫克曼两步骤方法，以及最大似然估计法在修正选择性样本偏误中的作用。他特别关注各方法的效率(参数估计量的方差)，提出了与伯克和雷(Berk & Ray, 1982)及其他许多研究者相同的问题。在误差服从二元正态分布时，以下三项重要因素会影响赫克曼估计量的表现：

1. 误差项 e 和 u 之间的相关系数 ρ
2. 两列解释变量 X 和 W 之间的相关性
3. 样本删截或选择的程度($z=1$ 的个案比例)

无论是尼尔森的研究，还是斯托增伯格和雷利的研究，都是将第三个因素固定，而使另外两个因素在各个模拟之间变化。在后者的研究中，样本的极大选择性(仅有10%的样本被选择)使OLS优于两步骤方法，因为在其他条件均等的情况下，它使后者的估计量在很大程度上失效。这是由于两步骤模型估计量的效率取决于用于修正样本选择性偏误的逆米尔斯比率 λ_i 与结果等式中的其他解释变量的相关程度。在杜安(Duan et al., 1984)等人的研究中，这项相关被表示为 R^2，即 λ_i 对 X_i 的测定回归系数。在因素三，即删截或样本选择的程度一定时，R^2 的大小取决于因素二。

假设 X 和 W 一致，则我们依赖probit的非线性来辨识结果模型，λ_i 是这些变量的非线性函数。但如果这些变量有限(即删截或样本选择很严重的情况

下)，则其线性函数会逐渐逼近非线性的 λ_i。换句话说，当样本选择性问题越加严重时，米尔斯比率会逐渐成为其构成变量的线性函数。因此，在 X 和 W 的相关性确定时，R^2 随样本选择性的极端化而增加。这会损害两步骤方法(及最大似然法)相对于 OLS(不适用米尔斯比率)的效率。

第 3 章讲到，在以下任何一个条件得到满足时，OLS 估计量是无偏的：(a)误差相关性 ρ 为 0；(b)λ_i 与结果等式中的解释变量不相关($R^2=0$)。

若两项条件均不满足，则 ρ 或 R^2 的增加(取决于因素二和因素三)会导致 OLS 系数偏误的增加(Nelson, 1984:193)。因此修正样本选择性偏误的需求更大。但是我们已经看到，任何非零的 R^2 都会导致两步骤模型(及最大似然估计)的效率的降低。因而广义看来这是一个两难的选择：两步骤估计量是一致的，而 OLS 估计则不是；但前者的参数估计量的方差更大。这同样建议我们应该谨慎地使用两步骤方法来估计 Tobit 模型，尤其是当两个等式中的解释变量相同的时候。表 2.2 说明相对于 OLS 和最大似然估计，两步骤 Tobit 模型的方差更大。

但斯托增伯格和雷利(Stolzenberg & Relles, 1990)关注这些估计量的效率而非准确性。由于 $\rho\neq0$ 且 $R^2\neq0$，这里不存在任何无偏估计，尽管两步骤估计量具有一致性。一致性是一种大样本特征，而斯托增伯格和雷利仅使用 500 个个案的样本，而且其删截率(未选择率)达到 90%，则其结果等式实际上只使用了 50 个个案。在这些条件下，我们当然不指望两步骤模型估计量有多么准确。但斯托增伯格和雷利论文中的表 4 还说明，只要 ρ 或者单个 x 与 w 之间的相关系数，两项之某一项超过 0.5，则两步骤模型的偏误小于 OLS 模型。在更大样本或不如此严重的删截数据情况下，模型的相对表现还会更加明显。

两步骤和最大似然方法的估计量的效率在 ρ 很低而 R^2 很高时最小。总的来说，最大似然估计的 β 和 ρ 比两步骤模型的估计量要更有效率。特别的，"OLS 偏误最大之条件，恰好也是最大似然估计对两步骤估计的优越性最大之条件(Nelson, 1984:195)。"尽管最大似然法比 OLS 估计量的方差要大，但其差异一般而言却较小(除非是 R^2 特别大的情况，如 >0.9)。因此，只要 OLS 有偏(且不一致)，则最大似然估计就优于 OLS 估计和两步骤估计。

评估研究中的样本选择模型

在劳动力市场项目的评估研究中，选择性样本模型受到了最多的批评(Fraker & Maynard, 1987; Lalonde, 1986)。这是因为此类项目很少使用随机分配，因而用以纠正选择性偏误的非实验估计量不能捕捉真实的(更准确地讲，是以随机分配为基础的)项目效应。这要求使用随机对照实验来评估此类项目(Ashenfelter & Card, 1985; Barnow, 1987)。在许多文章中，赫克曼都提到非试验方法的缺陷是由于选择性样本模型的错误使用造成的：这或者是由于对样本本质的错误认识，或者是因为纳入了不必要的或过于严格的假设。在再分析中，赫克曼、霍兹和戴伯斯(Heckman, Hotz & Dabos, 1987)以及赫克曼和霍兹(Heckman & Hotz, 1989)的文章都证明：这些非实验方法确实能给出与随机分配方法非常近似的结果。

由这部分观点看来，这一命题应关注以下两点：首先，评估研究及其他领域内的选择性偏误修正模型绝不止现有的这些(Heckman & Robb, 1986; Little & Rubin, 1987)；其次，合适的方法应由数据和我们观察到的现象所蕴含的社会过程来决定(例如受训者怎样被选择)。我们举两个例子。

若我们以考试成绩来测量两所学校 A 和 B 的效率。若数据为截面数据，则用最大似然法估计选择(学生进入哪类学校)和结果(进入某类学校后的考试成绩)等式是较合适的方法。若我们的数据给出两个时点的考试成绩，即学生在进入学校 A 或学校 B 之前的成绩以及在校内待过某段时间之后的成绩，则要修正选择性偏误，首先应写出第 i 位学生在第 t 次考试中的成绩为：

$$y_{it} = \beta X_{it} + \gamma z_i + u_{it}, u_{it} = \xi_i + v_{it}$$

而学生参与的选择等式为：

$$z_i = \alpha W_i + e_i, \ e_i = \xi_i + \varepsilon_i$$

此处 $z=1$ 表示学生进入学校 A，而 $z=0$ 则表示学生进入学校 B，对于 $t=1$ 的所有学生，$z=0$。由于共同因素 ξ 的存在，因而出现了样本选择性偏误。假设

v_{it}的均值为0，且方差为常数，且与ε和$v'_{it}(t' \neq t)$的所有值独立。则取首次差异，可以消去e和u之间的相关：

$$y_{i2} - y_{i1} = \beta(X_{i2} - X_{i1}) + \gamma z_i + u_{i2} - u_{i1}$$
$$= \beta(X_{i2} - X_{i1}) + \gamma z_i + v_{i2} - v_{i1}$$

则式中的误差项与其他所有解释变量独立，其均值为0，具有常数方差。因此，我们用两时点上解释变量的差异以及学校类型的虚拟变量对两时点的考试成绩差异进行回归。γ即表示在一类学校、而不是另一类学校的效应。很显然，我们可将ξ看作遗漏变量，它是各学生特有的、不随时间变化的特质，既影响其学校选择，亦影响其考试成绩。当我们取考试成绩随时间变化的值时，这项效应消失了。赫克曼和霍兹（Heckman & Hotz）基于对有观测的时点数目的考虑扩展了该模型。

另一处理贯穿本书的这类截面数据的办法为杜宾（Rubin，1977）的混合模型方法（参见 Land & McCall，1993）。杜宾在处理因变量的无填答问题时引入该方法，而由于无填答可以是多种样本选择性问题，因而该方法的应用范围理论上更广。杜宾用贝叶斯方法，在假设已选子样本和未选子样本之因变量分布参数的关系基础上，计算结果估计中可能的误差。换言之，研究者首先猜测或假设因变量在样本删截部分的分布，联合其对被选样本中因变量分布的认识，则对总体样本的因变量分布有所了解。该方法检测总体样本分布对未选择样本分布假设的敏感性。正如兰德和麦克科尔（Land & McCall，1993:302）所指出的，不同于选择偏误模型对选择过程的假设和拟合，相反的，混合模型方法对因变量的未观测分布进行假设。

这些假设被纳入混合模型，而解释变量的完全观测信息可用于形塑这些假设。根据每个假设，可计算一个贝叶斯预测概率区间（类似于置信区间），该类区间是广义上的一般置信区间。但假设所基于的预设信息也有可能并不清楚。在兰德和麦克科尔的例子中，假设未填答样本的均值与填答样本相同，但其方差更大。若我们有强烈的预设信息表明这些假设，或者结果对不同假设具有强烈的敏感性，则这些假设有很大的价值。但严重的问题在于：我们并不总是有

这样强烈的预设信息，而结果也并不总对不同假设敏感。这也说明，该模型只能用于测量未填答或选择性偏误的严重程度，但并不能对其进行修正。

删截模型和选择性样本模型的使用指南

我们将大略说明删截模型和选择性样本模型使用中应注意的事项。首先应考虑解决问题的适合模型。若数据还未收集，则我们应考虑最小化选择性样本问题。如在评估研究中，对控制组和反应组的参与者采用随机分配的原则。若数据已经收集完毕，则我们应从数据允许的分析方法中选择最合适的那个。若为历时性数据，则可使用上文介绍的赫克曼和霍兹(Heckman & Hotz, 1989)模型。若数据是截面的，则可使用第 2 章到第 4 章介绍的方法。

样本规模是一个重要的问题。本书讨论的多个估计量的优良特质都与大样本相关，如果样本很小，那么我们必须承认这类方法并不适合。因此，在仅使用本书方法才可解决的重要问题的研究中，研究设计必须保证有大样本。

若存在合适的大样本，且我们决定使用本书介绍的方法，则第一步应检验同方差和正态性。我们可以使用切希尔和艾利时(Chesher & Irish, 1987)的方法，或使用其他的替代性方法(Bera et al., 1984; Davidson & MacKinnon, 1984)。切希尔和艾利时详细介绍了在 Tobit 模型中怎样检验同方差和正态性，而我们可以将其扩展到选择性样本模型，但这时我们将检验二元正态性，因而其扩展形式更复杂。但在选择性样本模型中，我们可使用其方法对每一步骤——选择和结果——分开检验。在选择等式中检验正态性尤为重要。以 probit 独立估计该等式是使用切希尔和艾利时检验的最简单的方法。

若不满足同方差，则有必要对方差假设一个函数形式。类似的，若误差项不是正态分布，则我们应为其假设一个分布，或使用半参数模型。而当这些假设被满足时，对于删截回归模型，则 Tobit 估计量是可接受的。但对于选择性样本模型，我们应该先用 probit 得出逆米尔斯比率，再在结果等式中用解释变量对其进行回归。若 R^2 接近于 0，则结果等式可使用 OLS 估计，但如果并非如此，则我们应使用最大似然法估计两步骤模型。若最大似然模型中的 ρ 接近于

0，同样用 OLS 回归更合适。这是因为 OLS 估计中的 β 与最大似然估计值接近，但其方差更小（特别是在 R^2 很大时）。所以在选择性样本模型中，若 R^2 或 ρ 接近 0，则使用 OLS 方法；否则使用最大似然法。若可以使用最大似然法，则两步骤模型不具备任何优势。

最后一点是：怎样判断应使用删截回归（Tobit）还是选择性样本模型？马黛娜（Maddala，1992：54）提出了该问题，认为事实上 Tobit 模型在其大致所有的应用中都是不适合的，包括其最早的托宾的应用（Tobin，1958）。他认为 y 值本身的限制并不足以构成使用该模型的条件，相反，我们应该询问观测组形成的原因。若它是我们研究对象的某些决定带来（如不进行奢侈品消费）的，则删截模型并不适合。此处我们真正需要的是选择性样本模型——先单独拟合选择过程，再估计因变量的条件期望。而另一种情况则是 y 值的限制由外生性变量导致，如数据收集和记录（如本书开始的考试成绩例子），则毫无疑问删截模型是适合的。

当 y^* 为个体选择的函数时，使用删截模型还是选择性样本模型取决于我们对潜在变量 y^* 的本质的理解和解释。如我们假设 y^* 是渴望受教育的年限。我们假设对于完成最低年限后仍留在学校的人而言，其观测教育年限 y 即为 y^*。但对于那些在最低年限退出的人，其渴望年限小于或等于最低年限。若我们关注渴望年限，则应先考虑其在最低年限后是否留在学校的决定过程，删截模型这时就是适合的。而另一方面，奢侈品消费的潜在变量为消费愿望，它可以为负数。此时马黛娜可能是正确的，因为我们需要对是否消费和消费额度分别进行估计。这并不是方法论的问题，而是对问题和理论的理解的问题。

结论

本书介绍了当因变量存在删截、样本选择性和截断问题时常用的分析技术。我们将其与其他方法相联系，以举例说明基本方法怎样被扩展到不同的方面。在本书结尾，我们对模型的谨慎使用做出提醒。本书所讨论的估计量的优良特性只对大样本有效，且这些模型在违反正态性和同方差假设时，远

不如 OLS 模型表现得稳健。但选择性样本和删截问题在社会科学中非常普遍，本书用采纳这类技术的许多社会科学论文进行举例和检验。尽管我们需要谨慎使用这些方法，但毫无疑问它们能对大量问题提供有价值、有意义的解决办法。

附录 1 | 截断正态分布变量的期望值

该附录说明截断的正态分布随机变量之期望值的标准结果。令 u 为均值 0，标准差 σ 的正态分布随机变量，则当 u 自数值 m 被截断时，其期望值为：

A1 由上截断

$u \leqslant m$ 的概率为：

$$\Phi\left(\frac{m}{\sigma}\right)=\int^{\frac{m}{\sigma}} \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\mathrm{d}t$$

而 $u \leqslant m$ 时 u 的条件期望 $E(u \mid u \leqslant m)$ 为：

$$\frac{\phi(m/\sigma)}{\Phi(m/\sigma)}$$

A2 由下截断

$u > m$ 的概率为：

$$\int_{m/\sigma} \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)\mathrm{d}t = 1-\Phi\left(\frac{m}{\sigma}\right)$$

而 $u > m$ 时 u 的条件期望 $E(u \mid u > m)$ 为：

$$\frac{\phi(m/\sigma)}{1-\Phi(m/\sigma)}$$

由于正态分布的对称性，式子 A1 和 A2 也可由另一种方式写出，因而导致混淆。我们的观点是，正态分布的对称性此处意味着两点：首先，$1-\Phi(m/\sigma)=\Phi(-m/\sigma)$；其次，$\phi(m/\sigma)=\phi(-m/\sigma)$。例如我们定义 $m=c-X'_i\beta$，则 $E(u\mid u>c-X'_i\beta)$ 可写作：

$$\frac{\phi\left(\frac{c-X'_i\beta}{\sigma}\right)}{\Phi\left(\frac{X'_i\beta-c}{\sigma}\right)}=\frac{\phi\left(\frac{X'_i\beta-c}{\sigma}\right)}{\Phi\left(\frac{X'_i\beta-c}{\sigma}\right)}$$

后者则是我们在 Tobit 模型中使用的项。同样的，$E(u\mid u\leqslant c-X'_i\beta)$ 可以写作：

$$\frac{-\phi\left(\frac{c-X'_i\beta}{\sigma}\right)}{\Phi\left(\frac{c-X'_i\beta}{\sigma}\right)}=\frac{-\phi\left(\frac{X'_i\beta-c}{\sigma}\right)}{\Phi\left(\frac{c-X'_i\beta}{\sigma}\right)}=\frac{-\phi\left(\frac{X'_i\beta-c}{\sigma}\right)}{1-\Phi\left(\frac{X'_i\beta-c}{\sigma}\right)}$$

附录 2 | 切希尔和艾利时(Chesher & Irish)的正态性及异方差检验

使用切希尔和艾利时的检验需要估计四个矩残差，$\hat{e}^{(m)}$，m 等于 1 到 4。另 z_i 为表示是否删截的变量($z_i=0$ 表示删截)，并且 $k_i=X'_i\beta/\sigma$，且 $\lambda(k_i)$表示逆米尔斯比率。则 Tobit 模型的 4 个矩残差为：

$$\hat{e}^{(1)}=-(1-z_i)\lambda(k_i)+z_i\left(\frac{y}{\sigma}-k_i\right)$$

$$\hat{e}^{(2)}=(1-z_i)k_i\lambda(k_i)+z_i\left[\left(\frac{y}{\sigma}-k_i\right)^2-1\right]$$

$$\hat{e}^{(3)}=-(1-z_i)(2+k_i^2)\lambda(k_i)+z_i\left(\frac{y}{\sigma}-k_i\right)^3$$

$$\hat{e}^{(4)}=(1-z_i)(3k_i+k_i^3)\lambda(k_i)+z_i\left[\left(\frac{y}{\sigma}-k_i\right)^4-3\right]$$

实际中，我们用 β 和 σ 的估计值来计算这些数值。

为检验正态性，矩阵 R 的元素为：

$$\hat{e}^{(1)}X,\ \hat{e}^{(2)},\ \hat{e}^{(3)},\ \hat{e}^{(4)}$$

若 X 包含常数项，则 $\hat{e}^{(2)}$ 可被省略。在实际运用切希尔和艾利时的方法检验正态性时，我们对每一项观测都计算 $\hat{e}^{(m)}$ 的值，并以模型中的每一个变量 X_k(包括常数项)，$k=1, \cdots, K$，乘以 $\hat{e}^{(1)}$ 的值来构成一组新的变量。因而我们会得到有 $K+2$ 列的新矩阵 R。然后我们用一组向量一对 R 进行回归，得到解释平方和。拉格朗日乘数服从卡方分布，且在该例中自由度为 2。若拉格朗日乘数超过自由度为 2 的卡方值，则我们应拒绝正态性的零假设。

异方差检验遵循非常类似的方法。切希尔和艾利时同样给出了 probit 模型的 R 矩阵元素。这对我们非常有用,它可以为选择性样本模型的选择等式提供正态性检验。但有一点值得注意:当样本量很小时,该检验不可信。它只在样本量相对较大时适用。

注释

[1] 我们假设即使 y 有上限，也可因为估计需要将其忽略。

[2] 在格朗诺(Gronau，1974)和刘易斯(Lewis，1974)早期工作的基础上，赫克曼(Heckman，1976)最先提出这一结构并发展了这一两步骤方法。

[3] Tobit 为 Tobin probit 的缩写。

[4] 例如，离散变量 z 的期望值为：

$$\sum_i z_i \text{pr}(z = z_i)$$

而连续变量 z 的期望值为：

$$\int_{-\infty}^{\infty} y f(y) \mathrm{d}y$$

其中 f 表示密度函数。

[5] 关于一致性和渐进无偏性的区别，参见达利密斯(Dhrymes，1989:86—89)。

[6] 另一问题涉及对数似然函数的复杂性。例如，若我们定义的模型中包含非常复杂的多元积分，则最大似然估计法可能在事实上难以实施。

[7] 这四项偏导皆在假设 $c = 0$ 的情况下获得。但若该假设不成立，则除式 2.20c 外，其他几项偏导并不发生变化。对于式 2.20c，我们应考虑 x_j 的变化影响某观测值取非零阈值 c 的概率的局部效应，即在式中加入一项 $-\phi(z)\beta_j/\sigma$。

[8] 此处我们使用了微分的乘法法则。即在 $y = f(x)g(x)$ 时，y 对 x 的导数等于 $f(x)g'(x) + f'(x)g(x)$，其中$'$表示函数的导数。该分解式仅在 $c = 0$ 时成立，若该条件不满足，则我们应在式中加入一项，表示 x_j 的变化对观测值取非零阈值 c 的概率的影响。

[9] 由于霍诺汉和诺兰(Honohan & Nolan，1993)在其 Tobit 模型中并未报告误差项的方差估计，因而我们无法计算收入对拥有金融资产的概率的偏导数。

[10] 在第一章我们已经部分讨论过这一问题。斯托增伯格和雷利(Stolzenberg & Relles，1990，表 1)为我们列举了美国社会学评论(*American Sociological Review*)杂志中使用赫克曼技术修正选择性样本偏误的文章。

[11] 在因变量为收入的情况下，我们常令 y^* 为收入的转换形式，如其对数。

[12] 这是一个简化模型，它假设工作资源是充足的，因而使得个人是否获得工作完全取决于其自身选择。

[13] 此处我们讨论的是潜在变量 y^* 的误差项的异方差。应注意将其与第三章中讨论的赫克曼两步法里结果等式中的异方差问题相区别。

[14] 此时若使用本章稍后将讲到的一般化方法，仍可构建对数似然函数。李的方法的稳健性在用于多项选择模型时遭到了 Schmertmann 的质疑(1994)，尽管这项批评在仅仅使用二元选择时并不适宜。

[15] 计量经济学杂志，以及格拉格·邓肯(Greg Duncan)编辑的《未设定误差分布的连续与离散经济计量学》，都讨论了删截回归模型。

参考文献

ACHEN, C. H. (1982) *Interpreting and Using Regression*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—029. Beverly Hills, CA: Sage.

ALDRICH, J. H., and NELSON, F. D. (1984) *Linear Probability, Logit and Probit Models*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—045. Beverly Hills, CA: Sage.

ALLISON, P. D. (1984) Event History Analysis. Sage University Paper series on Quanntitative Applications in the Social Sciences, 07—046. Beverly Hills, CA: Sage.

ALLISON, P. D., and LONG, J. S. (1987) "Interuniversity mobility of academic scienntists." *American Sociological Review* 52:643—652.

AMEMIY A, T. (1979) "The estimation of simultaneous-equation Tobit model." *Internaational Economic Review* 20:169—181.

AMEMIY A, T. (1984) "Tobit models: A survey." Journal of Econometrics 24:3—61. ASHENFELTER, O., and CARD, D. (1985) "Using the longitudinal structure of earnillgs to estimate the effect of training programs." *Review of Economics and Statistics* 67:648—660.

BARNOW, B. S. (1987) "The impact of CETA programs on earnings: A review of the literature." *Journal of Human Resources* 22:157—193.

BARNOW, B. S., CAIN, G. G., and GOLDBERGER, A. S. (1980) "Issues in the analysis of selectivity bias." In E. W. Stromsdorfer & G. Farkas (Eds.), *Evaluation Studies Review Annual* (Vol. 5, pp. 43—59). Beverly Hills, CA: Sage.

BERA, A. K., JARQUE, C. M., and LEE, L. F. (1984) "Testing the normality assumption in limited dependent variable models." *International Economic Review* 25:563—578.

BERK, R. A. (1983) "An introduction to sample selection bias in sociological data." *American Sociological Review* 48:386—398.

BERK, R. A., and RAY, S. C. (1982) "Selection biases in sociological data." *Social Science Research* 11:352—398.

CHESHER, A., and IRISH, M. (1987) "Residual analysis in the grouped and censored normal linear model" *Journal of Econometrics* 34:33—61.

COLEMAN, J. S., HOFFER, T., and KILGORE, S. (1982) *High School Achievement: Public, Catholic and Other Private Schools Compared*. New York: Basic Books.

CRAGG, J. (1971) "Some statistical models for limited dependent variables with appliications to the demand for durable goods." *Econometrica* 39:829—844.

DAVIDSON, R., and MACKINNON, J. G. (1984) "Convenient specification tests for logit and probit models." *Journal of Econometrics* 25:241—262.

DEEGAN, J., Jr., and WHITE, K. J. (1976) "An analysis of nonpartisan election media expenditure decisions using limited dependent variable methods." *Social Science Research* 5:127—135.

DHRYMES, P. (1989) Topics in Advanced Econometrics. New York: Springer-Verlag. DUAN, N., MANNING, W. G., MORRIS, C. N., and NEWHOUSE, J. P. (1984) "Choossing

between the sample-selection model and the multi-part modeL" *Journal of Business and Economic Statistics* 2:283—289.

ELIASON, S. R. (1993) *Maximum Likelihood Estimation: Logic and Practice*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—096. Newbury Park, CA: Sage.

ENGLAND, P., FARKAS, G., KILBOURNE, B. S., and DOU, T. (1988) "Explaining occupational sex segregation and wages: Findings from a model with fixed effects." *American Sociological Review* 53:544—558.

FIN, T., and SCHMIDT, P. (1984) "A test of the Tobit specification against an alternative suggested by Cragg." *Review of Economics and Statistics* 66:174—177.

FRAKER, T., and MAYNARD, R. (1987) "Evaluating comparison group designs with employment-related programs." *Journal of Human Resources* 22:194—227.

GOLDBERGER, A. S. (1981) "Linear regression after selection." *Journal of Economettrics* 1S: 357—366.

GOLDBERGER, A. S. (1983) "Abnormal selection bias." In S. Karlin, T. Amemiya, & L. A. Goodman(Eds.), *Studies in Econometrics, Time Series and Multivariate Statisstics* (pp. 67—84). New York: Academic Press.

GREENE, W. H. (1981) "Sample selection bias as a specification error: A comment. "*Econometrica* 49:795—798.

GREENE, W. H. (1990) *Econometric Analysis*. New York: Macmillan.

GREENE, W. H. (1991) *LIMDEP User's Manual and Reference Guide*, Version 6.0. New York: Econometric Software.

GRONAU, R. (1974) "Wage comparisons—A selectivity bias." *Journal of Political Economy* 82:1119—1143.

HAGAN, J. (1989) Structural Criminology. New Brunswick, NJ: Rutgers University Press.

HAGAN, J., and PARKER, P. (1985) "White collar crime and punishment: The class structure and legal sanctioning of securities violations." *American Sociological Review* 50: 302—316.

HAUSMAN, J. A., and WISE, D. A. (1977) "Social experimentation, truncated distribuutions and efficient estimation." *Econometrica* 45:919—939.

HECKMAN, J. J. (1974) "Shadow prices, market wages and labour supply." *economettrica* 42:679—694.

HECKMAN, J. J. (1976) "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models." *Annals of Economic and Social Measurement* 5:475—492.

HECKMAN, J. J. (1979) "Sample selection bias as a specification error." *Econometrica* 47:153—161.

HECKMAN, J. J. (1992) "Selection bias and self-selection." In J. Eatwell, M. Milgate, & P. Newman(Eds.), *The New Palgrave Econometrics* (pp. 201—224). London: Macmillan.

HECKMAN, J. J., and HOTZ, V. J. (1989) "Choosing among alternative nonexperimental meth-

ods for estimating the impact of social programs: The case of manpower training." *Journal of the American Statistical Association* 84(408):862—874.

HECKMAN, J. J., HOTZ, V. J., and DABOS, M. (1987) "Do we need experimental data to evaluate the impact of manpower training on earnings?" *Evaluation Review* 11:395—427.

HECKMAN. J. J., and ROBB, R. (1986) "Alternative identifying assumptions in econometric models of selection bias." *Advances in Econometrics* 5:243—287.

HONOHAN, P., and NOLAN, B. (1993) *The Financial Assets of Households in Ireland* Dublin: The Economic and Social Research Institute, General Research Series Paper 162.

JOHNSTON, J. (1972) *Econometric Methods* (2nd ed.). Tokyo: McGraw-Hill.

KALBFLEISCH, J. D., and PRENTICE, R. L. (1980) *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

KARLIN, S., and TAYLOR, H. M. (1975) *A First Course in Stochastic Processes*. Volume 1 (2nd ed.). New York: Academic Press.

KMENTA, J. (1971) *Elements of Econometrics*. New York: Macmillan.

LALONDE, R. (1986) "Evaluating the econometric evaluations of training programs with experimental data." *American Economic Review* 76:604—620.

LAND, K. C., and MCCALL, P. L. (1993) "Estimating the effect of nonignorable nonresponse in sample surveys." *Sociological Methods & Research* 21:291—316.

LEE, L. F. (1983) "Generalized econometric models with selectivity." *Econometrica* 51:507—512.

LEE, L. F. (1994) "Semiparametric two stage estimation of sample selection models subject to Tobit-type selection rules." *Journal of Econometrics* 61:305—344.

LEE, L. F., and MADDALA, G. S. (1985) "The common structure of tests for selectivity bias, serial correlation, heteroscedasticity and non-normality in the Tobit model." *International Economic Review* 26:1—20.

LEWIS, H. G. (1974) "Comments on selectivity bias in wage comparisons." *Journal of Political Economy* 82:1145—1156.

LEWIS-BECK, M. S. (1980) *Applied Regression: An Introduction*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07—022. Beverly Hills, CA: Sage.

LITTLE, R. J. (1985) "A note about models for selectivity bias." *Econometrica* 53:1469—1474.

LITTLE, R. J., and RUBIN, D. B. (1987) *Statistical Analysis With Missing Data*. New York: John Wiley.

MADDALA, G. S. (1983) *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

MADDALA, G. S. (1992) "Censored data models." In J. Eatwell, M. Milgate, & P. Newman (Eds.), *The New Palgrave Econometrics* (pp. 54—57). London: Macmillan.

MCDONALD, J. F., and MOFFIT, R. F. (1980) "The uses of Tobit analysis." *Review of Economics and Statistics* 62:318—321.

MCKELVEY, R. D., and ZAVOINA, W. (1976) "A statistical model for the analysis of ordinal level dependent variables." *Journal of Mathematical Sociology* 4:103—120.

NELSON, F. D. (1984) "Efficiency of the two-step estimator for models with endogenous sample

selection. " *Journal of Econometrics* 24:181—196.

NEWEY, W. K., POWELL, J. L., and WALKER, J. R. (1990) "Semiparametricestimation of selection models: Some empirical results. " *American Economic Review*, AEA Papers and Proceedings 80:324—328.

OLSEN, R. J. (1978) "Comment on the uniqueness of the maximum likelihood estimator for the Tobit model. " *Econometrica* 46:1211—1215.

OLSEN, R. J. (1980) "A least squares correction for selectivity bias. " *Econometrica* 48:1815—1820.

OLSEN, R. J. (1982) "Distributional tests for selectivity bias and a more robust likelihood estimator. " *International Economic Review* 23:223—240.

PETERSON, R., and HAGAN, J. (1984) "Changing conceptions of race: Towards an account of anomalous findings of sentencing research. " *American Sociological Review* 49:56—70.

POWELL, J. L. (1984) "Least absolute deviations estimation for the censored regression modeL" *Journal of Econometrics* 25:303—325.

RUBIN, D. B. (1977) "Formalizing subjective notions about the effect of nonrespondents in sample surveys. " *Journal of the American Statistical Association 72*(359):538—543.

SANDERS, J. M., and NEE, V. (1987) "Limits of ethnic solidarity in the enclave economy. " *American Sociological Review* 52:745—773.

SCHMERTMANN, C. P. (1994) "Selectivity bias correction methods in polychotomous sample selection models. " *Journal of Econometrics* 35:101—132.

STEW ART, M. (1983) "On least-squares estimation when the dependent variable is grouped. " *Review of Economic Studies* 50:141—149.

STOLZENBERG, R. M., and RELLES, D. A. (1990) "Theory testing in a world of constrained research design. " *Sociological Methods & Research* 18:395—415.

TIENDA, M., SMITH, S. A., and ORTIZ, V. (1987) "Industrial restructuring, gender segregation and sex differences in earnings. " *American Sociological Review* 52:195—210.

TOBIN, J. (1958) "Estimation of relationships for limited dependent variables. " *Economettrica* 26:24—36.

WALTON, J., and RAGIN, C. (1990) "Global and national sources of political protest: Third World responses to the debt crisis. " *American Sociological Review* 55:876—890.

译名对照表

accelerated failure time model	加速失效模型
asymptotic property	渐进性质
asymptotically unbiased estimation	渐进无偏估计
censored data	删截数据
covariance matrix	协方差矩阵
double truncated	双重截断
evaluation research	评估研究
explicit selection	外在选择
globally concave function	广义凹函数
heteroscedasticity	异方差
homoscedasticity	同方差
incidental selection	内在选择
information matrix	信息矩阵
inverse Mill's ratio	逆米尔斯比率/风险率
joint probability	联合概率
latent variables	潜在变量
left truncated	左截断
likelihood function	似然函数
log-likelihood	对数似然值
marginal probability	边缘概率
maximum likelihood estimation	最大似然估计
mild regularity conditions	非极端正则条件
model identification	模型辨识
moment residuals	矩残差
predictive Bayesian probability interval	贝叶斯预测概率区间
right truncated	右截断
robustness	稳健性
sample selected data	选择性样本数据
threshold	阈值
truncated data	截断数据

分位数回归模型

作 者 简 介

郝令昕

郝令昕(Lingxin Hao),1990 年获芝加哥大学社会学博士,目前是约翰·霍普金斯大学的社会学教授,Russell Sage 基金会 2002 年至 2003 年的访问学者。她的专业领域包括家庭与公共政策、社会不平等、移民、定量研究方法和高级统计。她研究的焦点是美国家庭,强调除个体和家庭因素之外的结构、制度和情境的作用,并通过高级统计方法和大规模全国调查数据,来检验从社会学和经济学理论中推导出来的研究假设。其著作发表在各类杂志上,包括《社会学方法论》、《社会学方法与研究》、《定性与定量研究》、《美国社会学杂志》、《社会力、教育社会学、社会科学研究》和《国际移民研究评论》。

丹尼尔·奈曼

丹尼尔·奈曼(Daniel Q. Naiman),1982 年获伊利诺伊大学厄巴纳－尚佩恩分校数学博士,目前是约翰·霍普金斯大学应用数学与统计系教授兼系主任。1997 年被选为数学统计研究所会员,并于 2005 年成为坎特伯雷大学的厄斯金研究员(Erskine fellow)。他的数学研究工作主要集中在进行多元检验的几何与计算方法。他应用统计技术与他人合作的领域十分广泛:生物信息学、经济计量学、环境健康、基因、水文学和微生物学等。其作品出现在各类杂志上,包括《统计学年报》、《生物信息学》、《生物统计学》、《人类遗传》、《多元分析杂志》、《美国统计学会杂志和自然科学》。

译 者 简 介

肖东亮

肖东亮,香港科技大学社会科学部硕士研究生,其研究领域为社会分层与流动、劳动力市场和经济社会学。

序

40 年来，经典的线性回归模型一直是定量社会科学研究者方法论中重要的组成部分。目前已有的关于定量方法应用的书籍，涉及许多关于线性回归的各种延伸话题，例如 logit、probit、事件史(event history)、广义线性(generalized linear)、广义非参数模型(generalized nonparametric models)和处理删截、样本选择、截断和缺失数据的线性回归模型(linear-regression models)；此外，还包括许多其他相关的模型，例如方差分析、协方差分析、因果模型、对数线性模型、多重比较和时间序列分析。

经典回归的主旨是估计因变量在解释变量特定值域之上的条件均值。当回归假设成立时，这一方法是有效的；但当出现非标准情况时，它就会失效。(关于线性回归假设的详细讨论，见本丛书第一辑“理解回归假设”，William Berry 著)其中两个是正态性假设和方差齐性假设。通常的社会科学数据无法满足这两个关键的假设。例如，(条件)收入分布一般不是正态的，并且首席执行官的年度分红分布随着公司规模的增大而上升，这意味着存在异方差性问题。这正是分位数回归可以处理的问题，因为它放松了这些假设。另外，分位数回归为研究者提供了一个(无法从经典回归中获得的)新视角，研究解释变量对响应变量分布中位置、尺度和形状的效应。

分位数回归的思想并不新颖，事实上它起源于 1760 年，当时一个游历学者克罗地亚基督徒 Rudjer Josip Boscovich——他拥有许多头衔：物理学家、天文学家、外交官、哲学家、诗人和数学家——来到伦敦讲授他尚未成熟的中位数回归方法。然而，这一回归方法计算的复杂性直到最近依然是一大挑战。由于今日快速的计算功能和统计软件的广泛应用(如可执行分位数回归程序的 R、SAS 和 Stata)，使得拟合分位数回归模型变得更加容易。但是，至今我们仍未提供任何关于分位数回归是什么的介绍。在本书中，Hao 和 Naiman 提出了分位数和分位数函数的概念，并阐述了分位数回归模型，讨论了它们的估计和推断方法，并通过具体例子演示了对分位数回归估计值(是否转换)的解释。同时，他们也提供了应用分位数回归分析美国 1991 年和 2001 年收入不平等的完整例子，以此确定这一方法的思想和步骤。本书填补了丛书的空白并且有助于社会科学研究者更加熟悉分位数回归。

廖福挺

第 1 章 | 引言

回归分析的目的在于揭示因变量和自变量的关系。在实际的应用中，自变量并不能精确地估计因变量。相反，与每个自变量的特定值相对应的响应变量是一个随机变量。因此，我们常常使用集中趋势的测量方法，来概括自变量特定值域下的因变量变化情况，主要包括均值（mean）、中位数（median）和众数（mode）。

传统的回归分析主要关注均值，即采用因变量条件均值（conditional mean）的函数来描述自变量每一特定数值下的因变量均值，从而揭示自变量与因变量的关系。模型化和拟合条件均值函数（conditional-mean function）是回归模型法大族谱中的核心思想，具体包括常见的简易线性回归模型、多元回归、加权最小平方数下的异方差误差模型（heteroscedastic error）和非线性回归模型。

条件均值模型（conditional-mean models）具有以下优点：在理想的条件下，它们可以为我们提供关于自变量和因变量分布关系的完整的和参数的描述。另外，采用条件均值模型可获得具有优越统计特性的估计量（最小二乘法和最大似然法），它更容易计算，并且更容易解释。这种模型通过不同的方式被推广，从而适用于误差具有异方差性的情况，因此，对于特定的自变量，因变量条件均值和条件单位（conditional scale）的模型化可以同时进行。

条件均值模型被广泛应用在社会科学中，尤其在过去的半个多世纪里，使用最小二乘法及其衍化方法对连续型因变量和自变量的关系进行回归建模被认为是现代重要的统计工具。最近，分析二分因变量的 logistic 和 probit 模型、分析计数因变量的泊松回归模型在社会科学研究中的重要性不断提高。这些方法并没有超出条件均值模型的框架（conditional-mean modeling framework）。

当定量社会科学的研究者已经应用更高级的分析方法来放宽条件均值框架下的一些建模假设时，这个框架本身却很少被质疑。

条件均值框架存在先天的局限性。首先，当归纳自变量特定数值下的响应变量情况时，这个条件均值模型并不能轻易地扩展到非中心位置(noncentral locations)，而非中心位置往往正是社会科学研究的兴趣所在。例如，关于经济不平等和流动的研究对穷人(低尾)和富人(上尾)的情况有浓厚的兴趣。教育研究者会设法在既定的成绩水平下去理解和减少群体差异(如三层次参照标准：基础、熟练和高级)。这样，对中心位置的强调，长期阻碍了学者采用恰当的技术来研究有关因变量非中心位置的课题。而采用条件均值模型来分析以上问题是没有效率的，甚至会偏离研究重点。

其次，这些模型的假设在现实生活中并不总会得到满足。特别是方差齐性假设经常落空。另外仅仅关注集中趋势会忽视关于因变量分布的有用信息。并且，社会现象中通常会出现重尾分布，从而导致离群值优势。正因为条件均值深受离群值的干扰，所以它对中心位置的测量是不恰当和具有误导性的。

最后，一直以来对中心位置的关注转移了学者对因变量整体分布性质的注意力。我们需要跳出预测变量的位置和数值范围对因变量的效应这一框架，进而探讨预测变量的变化会如何影响因变量分布的基本形状。例如，许多社会科学研究关注社会分层和不平等，这一领域要求深入分析因变量的分布特征。对分布特征的描绘包括中心位置、数值范围(scale)、偏态和其他高阶特性，而不仅仅是中心位置。因此，采用条件均值模型来表述因变量分布与自变量的关系是具有先天性缺陷的。关于不平等主题的例子包括工资、收入和财富等经济不平等；在学业成绩上的教育不平等；在身高、体重、疾病发生概率、毒品上瘾、医疗和预期寿命上的健康不平等和由于社会政策而导致的生活质量的不平等。这些课题通常采用条件均值框架进行分析，从而忽略了其他更重要的分布特征。

条件均值模型的替代方法可以追溯到 18 世纪中期。这一方法被称为条件中位数模型，或简称中位数回归(median regression)。它解决了一些上面提出的关于集中趋势测量方法的选择问题。这种方法用最小绝对距离估计(least-absolute-distance estimation)代替最小二乘估计(least-squares estimation)。最

小二乘估计不需要大功率的计算机便可轻松实现，然而最小绝对距离估计必须借助强大的计算机力量。所以，直到20世纪70年代后期，当计算机技术融合了如线性优化等算法系统时，采用最小绝对距离估计的中位数回归模型才变得实用。

中位数回归模型可以实现与条件均值回归模型同样的目标：表述因变量的中心位置与一组协变量的关系。然而，当因变量的分布是高度偏态时，均值在解释的时候就会受到质疑，而中位数依然保有大量信息。因此，条件中位数模型具有更大的应用潜力。

中位数是一个特殊的分位数，它表示一种分布的中心位置。中位数回归是分位数回归的一种特殊情况，在这里第0.5分位数被模型化为一个关于协变量的函数。一般地说，其他分位数则可以用来描述一种分布的非中心位置。分位数概念可归纳为一些特定的名称，如四分位数、五分位数、十分位数和百分位数。第p个百分位数表示因变量的数值低于这一百分位数的个案数占总体的p%。因此，分位数可以指定分布中的任何一个位置。例如，有2.5%的个案数值低于第0.025分位数。

凯恩克(Koenker)和巴西特(Bassett)在1978年引入分位数回归，将条件分位数模型化为预测变量的函数。分位数回归模型是线性回归模型的自然扩展。随着协变量的变化，线性回归模型描述了因变量条件均值的变化，而分位数回归模型则强调条件分位数的变化。由于所有分位数都是可用的，所以对任何预先决定的分布位置进行建模都将是可能的。因而，研究人员可以选择适合他们特定研究议题的分位数进行分析。贫穷研究关心低收入人群，例如，在2000年11.3%的社会底层生活在贫穷状态中(U. S. Census Bureau, 2001)。税收政策研究则关注富人，例如，最富有的4%的人口(Shapiro & Friedman, 2001)。条件分位数模型为集中研究人口中的特定人群提供了灵活性，而条件均值模型则做不到。

由于多元分位数可被模型化，所以我们可以更加全面地理解因变量的分布是如何受到预测变量的影响的，包括形状变化(shape shift)等信息。一组间距相同的条件分位数(例如总体中的每5%或每1%)可以描绘除中心位置外的条

件分布的形状。这种模型化形状变化的能力是社会不平等研究领域在方法论上的一次飞跃。按照惯例，以往的不平等研究并不是建立在模型基础上的，这些方法包括洛伦兹曲线（Lorenz curve）、基尼系数（the Gini coefficient）、泰尔熵标准（Theil's measure of entropy）、方差系数和对数转换分布的标准差。

通过建立在线性优化基础上的算法系统，最小化关于距离的广义测量方法，便可以轻松地建立分位数回归模型。因此，分位数回归目前是研究者的实用工具。社会科学家所熟悉的软件包则提供了简单易懂的命令来拟合分位数回归模型。

在凯恩克（Koenker）和巴西特（Bassett）首次引入分位数回归的 15 年后，有关分位数回归的实际应用开始迅速普及。实证研究者通过分位数回归来检验预测变量对因变量分布的影响。由经济学家（Buchinsky, 1994; Chamberlain, 1994）完成的两篇早期实证研究论文，为我们提供了如何将分位数回归应用到工资研究中的实际例子。借助分位数回归，他们全面分析了工资的条件分布，发现教育和工作经验的回报以及工会成员身份的效应在不同的工资分位点上是不同的。采用分位数回归分析工资的例子不断增加，并且扩展至另外一些话题，如工资分布的变化（Machado & Mata, 2005; Melly, 2005），特定行业内的工资分布（Budd & McCall, 2001），白人与少数族裔（Chay & Honore, 1998）以及男性与女性（Fortin & Lemieux, 1998）的工资差距，受教育水平和工资不平等（Lemieux, 2006）以及收入的代际转移（Eide & Showalter, 1999）。分位数回归同样应用于分析学校的教育质量（Bedi & Edwards, 2002; Eide, Showalter & Sims, 2002）以及人口特征对婴儿出生体重的影响（Abreveya, 2001）。分位数回归还延伸至其他领域，特别是社会学（Hao, 2005, 2006a, 2006b）、生态学和环境科学（Cade, Terrell & Schroeder, 1999; Scharf, Juanes & Sutherland, 1989），还有医学和公共卫生等领域（Austin et al., 2005; Wei et al., 2006）。

本书旨在向那些对分布形状和位置的建模方法有着浓厚兴趣的社会科学家们介绍分位数回归模型。此外，本书同样适合那些关注线性回归模型易受偏态分布和离群值影响这一问题的读者们。该书的写作主要建立在凯恩克（Koenker）及其同事们的奠基性著作上（例如：Koenker, 1994; Koenker, 2005;

Koenker & Bassett，1978；Koenker & d'Orey，1987；Koenker & Hallock，2001；Koenker & Machado，1999)并作出了两大贡献。在分位数回归估计值的基础上,我们发展了基于条件分位数上的形状变化的测量方法。这些测量方法为我们提供了关于协变量如何影响因变量的分布形状这一研究问题的直接答案。另外,不平等研究常常对右偏的因变量分布采用对数转换方式,来获得更好的模型拟合效应,尽管在这种情况下"不平等"代表着初始数值的分布。因此,我们发展出一套方法,从对数单位系数中计算协变量对条件分位数函数的位置和形状的绝对值效应。

从我们的研究经验中知道,这本书是为从事实证研究的学者而编写的。我们采用社会科学家熟悉的语言和步骤进行教学,具体包括定义清晰的术语、简化的方程式、插图、实证数据的图表和社会科学家熟悉的统计软件的计算编码。贯穿全书,我们从自己的家庭收入研究中提取实际例子进行讲述。为了更好地介绍分位数回归,我们使用简化的模型设定,在这里,不管是初始单位还是对数转换的因变量,其条件分位数函数对于协变量而言都是线性的和可加的。正如在线性回归中,我们介绍的方法可以轻松地应用在更加复杂的模型设定中,例如交互项和协变量的多项式或样条函数(spline function)。

本书内容组织如下:第 2 章从两个方面定义分位数和分位数函数——运用分布函数和解决最小化问题。相对于分布矩阵(如均值、标准差),本章还提出测量分布位置和形状的分位差方法(quantile-based measures)。第 3 章比较了线性回归模型和分位数回归模型(QRM)的基本原理,包括模型建立、估计量和特性。通过特定的分位数参数来建构多条分位数回归方程是分位数回归模型的独特性质。我们将展示如何运用最小距离原则(minimum distance principle)来拟合分位数回归方程。QRM 假设分布具有单调同变性(monotonic equivariance principle)和稳健性等特性,这些特性可为我们提供灵活稳健的估计,此外,QRM 还具有其他线性回归模型所不具备的性质。在第 4 章里,我们讨论了分位数回归模型的推论方法。除了介绍分位数回归系数的渐近推论(asymptotic inference)外,本章还强调自举法(bootstrap method)的实用性和可行性。另外,相对于线性回归模型,我们还简短地讨论了分位数回归模型的拟合优度。第 5

章提出了多种解释分位数回归估计值的方法。本章超越了协变量对特定条件分位数(如中位数或其他非中心分位数)效应的传统检验,主要关注对分布的理解(distributional interpretation)。它阐述了对分位数回归估计值的图像化解释和从分位数回归估计值中对形状变化的定量测量,包括位置转移、尺度变化(scale shift)和偏态变化等。第 6 章讨论与单调转换因变量相关的话题。我们发展了两种方法,从对数单位系数中获得协变量对条件分位数函数的位置和形状的绝对值效应。第 7 章讲述了本书介绍并加以发展的技术的系统运用。在本章中,我们分析了美国在 1991 年至 2000 年持续并扩大的收入不平等的原因。最后,附录提供了执行第 7 章所描述的分析任务的 Stata 命令。

第 2 章 | 分位数和分位数函数

描述并比较总体的分布特征，是社会科学的本质。描述分布最简单和最常见的方法，莫过于寻找代表中心位置的平均值和揭示离散程度的标准差。然而，将注意力仅仅局限于平均值和标准差，无疑会让我们忽视其他有助于深入挖掘分布特征的重要特性。对于许多学者而言，他们感兴趣的总体多为偏态分布，因此均值和标准差并不是测量位置和形状的最佳方法。为了描绘非对称分布的位置和形状特征，本章用分布函数(CDF)的方法介绍分位数、分位数函数及其特性，并且发展出测量分布位置和形状的分位差方法，最后将分位数重新定义为最小化问题的解决方法。

分布函数、分位数和分位数函数

我们可通过分布函数描绘一个随机变量 Y 的分布。分布函数即在给定的函数 F_Y 中，对于每一个 y 值，当 $Y \leqslant y$ 时在总体中所占的比例。图 2.1 呈现了标准正态分布的分布函数。分布函数可用作计算 y 值在任意区间占总体的比例。由图 2.1 可知，$F_Y(0) = 0.5$ 和 $F_Y(1.28) = 0.9$。我们可以通过这一函数计算所有其他关于 Y 的概率。特别有 $P[Y > y] = 1 - F_y(y)$（例如，在图 2.1 中，$P[Y > 1.28] = 1 - F_y(1.28) = 1 - 0.9 = 0.1$）和 $P[a < Y \leqslant b] = F_Y(b) - F_Y(a)$)（例如，在图 2.1 中，$P[0 \leqslant Y \leqslant 1.28] = F_Y(1.28) - F_Y(0) = 0.40$）。分布函数最重要的两大特性是单调性（例如，当 $y_1 \leqslant y_2$ 时，$F(y_1) \leqslant F(y_2)$）和它的极限 $\lim_{y \to -\infty} F(y) = 0$ 和 $\lim_{y \to +\infty} F(y) = 1$。对于一个连续型随机变量 Y，我们同样可用概率密度函数来表示它的分布，对于所有 a 和 b 的取值，都有

$P[a \leqslant Y \leqslant b] = \int_{y=a}^{b} f_Y dy$。

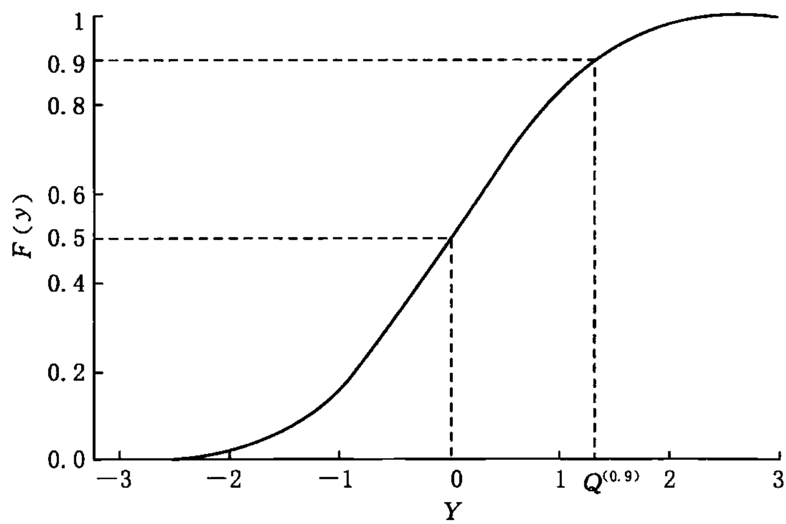


图 2.1　标准正态分布的累积密度函数

让我们回到通过位置和离散并不足以充分描述一个分布的话题。假如我们知道白人家庭的平均收入(W)比黑人家庭的平均收入(B)多出 20500 美元。这可以简单描述成形状不变的分布图在位置上的移动(见图 2.2a 中对应的密度函数),因此这两种分布的关系可表示为 $F^B(y) = F^W(y - 20500)$。但事实上,这两种分布的差异同时体现在位置和形状上(见图 2.2b 中对应的密度函数),所以对于常数 a 和 $c(a > 0)$,两种分布之间的关系可归纳为 $F^B(y) = F^W(ay - c)$。这就是当 y 的均值和方差在总体 W 和 B 之间都不相同时出现的情况。对位置和尺度(location and scale)的测量方法,如均值和标准差,或者中位数和四分位距,有助于我们比较两种分布的 Y 属性。

分布越不对称时,需要越复杂的分析方法。对分位数和分位数函数的考虑可为我们提供一系列丰富的分析方法。下面我们继续讨论分布函数 F,对于某些总体特征而言,该分布的第 p 分位数可表示为 $Q^{(p)}(F)$ [或者当被讨论的总体是已知时,可简化为 $Q^{(p)}$], $Q^{(p)}(F)$ 则代表分布函数在 p 点上的反函数(inverse)的值,即存在一个值 y,使得 $F(y) = p$。所以,处于 $Q^{(p)}$ 值之下的比例为 $p\%$。例如,在标准正态分布的例子中(见图 2.1),因为 $F(1.28) = 0.9$,所以

$Q^{(0.9)}=1.28$，那就是在值 1.28 之下的比例为 0.9 或 90%。

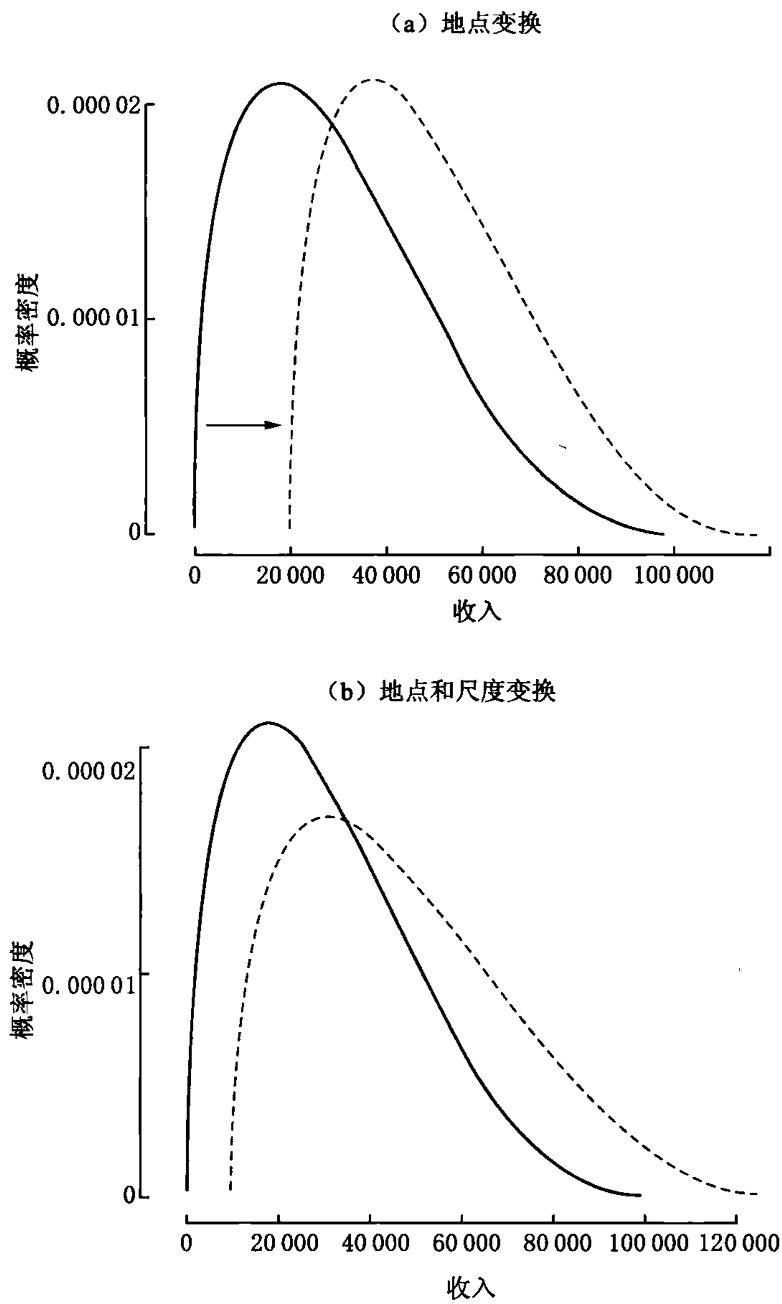


图 2.2 位置移动、位置和尺度变化(假设数据)

类似于总体的分布函数，我们考虑对应于一个样本的经验或样本分布函数(empirical or sample cdf)。对于一个包含值 $y_1, y_2, \cdots, y_n$ 的样本，经验分布函

数表示样本值小于或等于任意值 y 所占的比例。经验分布函数 $\hat{F}$ 的正式的定义为：

$$\hat{F}(y) = \text{样本值小于或等于 } y \text{ 值所占的比例}$$

例如，考虑一个包含 20 户家庭收入情况的样本（单位：美元），3000、3500、4000、4300、4600、5000、5000、5000、8000、9000、10000、11000、12000、15000、17000、20000、32000、38000、56000 和 84000。由于 8 户家庭的收入在5000 美元以下，所以我们有 $F(5000) = 8/20$。这一经验分布函数图如图 2.3 所示，包括一个上涨部分和几个平坦部分。例如，在 5000 美元这一点，上涨部分达 3/20，这表明在该样本中，5000 美元出现了 3 次；而平坦部分则出现在56000 美元和 80000 美元之间，这表明在这两点之间不存在其他样本值。因为经验分布函数存在平坦部分，所以在某些值上存在多个反函数。例如，如图 2.3 所示，在 56000 和 80000 之间的 $Q^{(0.975)}$ 是一个存在多种选择的连续统。因此，当我们采用分位数和分位数函数来分析一般性分布时，需要留意以下定义：

定义：一条分布函数下的第 p 分位数 $Q^{(p)}$ 是一组 y 值中的最小值，从而使 $F(y) \geqslant p$。函数 $Q^{(p)}$（作为 p 的函数）正是 F 的分位数函数。

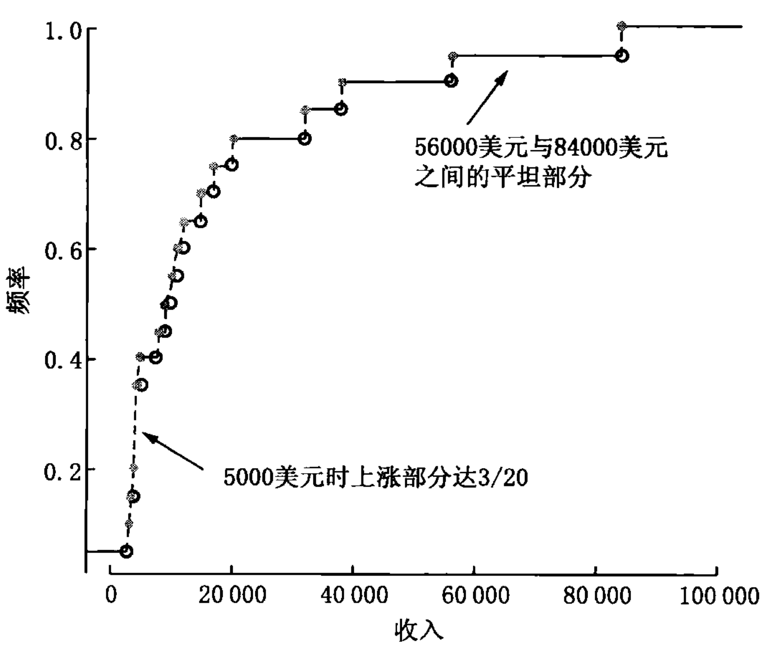


图 2.3　具有跳跃和水平部分的累积密度函数

图 2.4 展示了分位数函数和与之相应的分布函数。由此可以观察到:分位数函数是一条从底端开始的单调非递减的连续性函数。

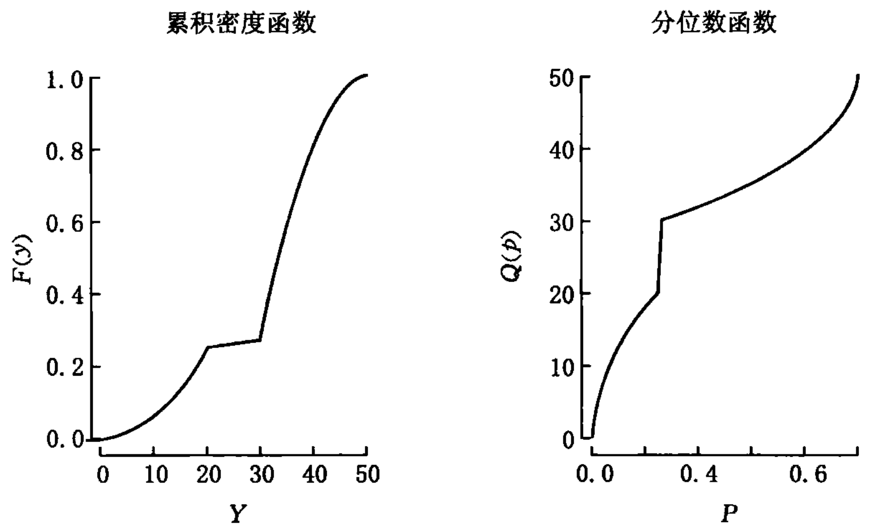


图 2.4　累积密度函数及其相应的分位数函数

作为特殊的例子,我们讨论样本分位数(sample quantiles),它可用作估计抽样分布(sampled distribution)的分位数。

定义:给定一个样本 y_1, y_2, …, y_n,我们将第 p 样本分位数$\hat{Q}^{(p)}$定义为相应的经验分布函数 $\hat{F}$ 的第 p 分位数,即 $\hat{Q}^{(p)} = \hat{Q}^{(p)}(F)$。与之相应的分位数函数,就表示为样本分位数函数。

样本分位数与顺序统计量(order statistics)密切相关。假定样本 y_1, y_2, …, y_n,我们按其大小从低到高排列,则表示为 $y_{(1)}$, …, $y_{(n)}$,有 $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$。有些数值如果出现多次,它们会被重复。顺序统计量和样本分位数的关系可简单描述如下:对于大小为 n 的一个样本,第(k/n)样本分位数由 $y_{(k)}$ 确定。例如,在以上的 20 户家庭收入数据中,第(4/20)的样本分位数,即第 20 百分位数,等于 $\hat{Q}^{(0.2)} = y_{(4)} = 4300$。

样本分位数的抽样分布

样本分位数在大样本中会如何表现,是需要重点关注的。对于一个从某分

布中抽取的大样本 $y_{(1)}$，…，$y_{(n)}$，该分布的分位数函数为$\hat{Q}^{(p)}$，概率密度函数为 $f=F$，$\hat{Q}^{(p)}$ 的分布接近均值为 $Q^{(p)}$、方差为 $\frac{p(1-p)}{n}\cdot\frac{1}{f(Q^{(p)})^2}$ 的正态分布。特别的，这一样本分布的方差完全由在分位点上估计而来的概率密度决定。这种对分位点上的密度的依赖有简单直观的解释：如果分位数附近有较多数据点（更高的密度），那么样本分位数更稳定；相反的，如果分位数附近有较少数据点（更低的密度），那么样本分位数较不稳定。

为了估计分位数抽样的变异性，我们可以利用以上的方差近似值，但这需要事先估计未知的概率密度函数。图 2.5 给我们展示了一种标准的估计方法，函数$\hat{Q}^{(p)}$在点 p 的切线斜率是分位数函数在 p 点上的导数，同样的，有密度函数的倒数：$\frac{d}{dp}Q^{(p)}=1/f(Q^{(p)})$。这一项式接近点 $(p-h,\ \hat{Q}^{(p-h)})$ 和 $(p+h,\ \hat{Q}^{(p+h)})$ 的割线斜率 $\frac{1}{2h}(\hat{Q}^{(p+h)}-\hat{Q}^{(p-h)})$，尤其当 h 为极小值时。

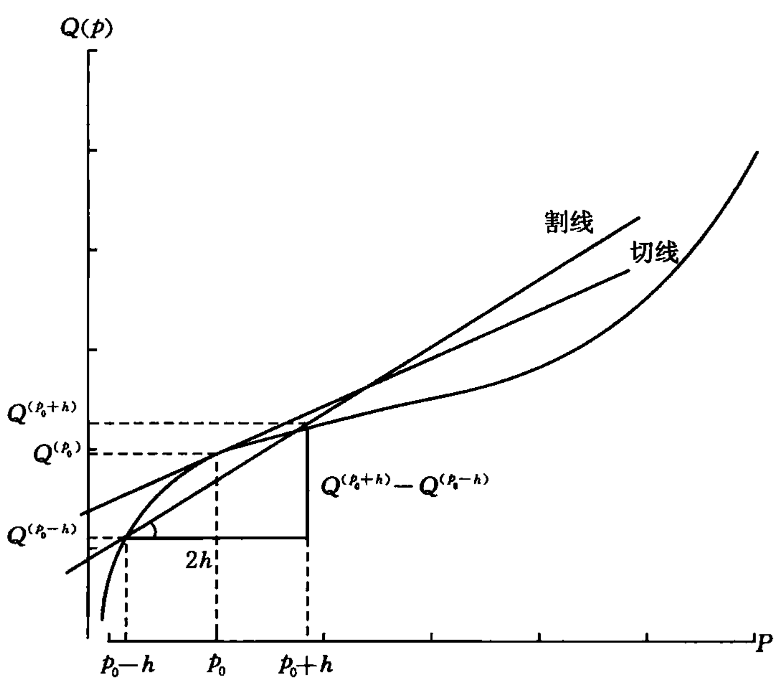


图 2.5 如何估计分位数函数的斜率的图示

注：函数在 p_0 点的导数（切线的斜率）约等于割线的斜率。

位置和形状的分位差测量方法

社会科学家们熟悉关于中心位置的分位差测量方法，即是中位数（如 50%分位数）而不是均值（密度函数的第一阶矩）被用作指示偏态分布的中心。

通过基于分位数的位置允许我们研究更多非分布中心的位置概念。具体的，我们可以由此检验特定子样本中的低尾位置（如 10%分位数）或者上尾位置（如 90%分位数）。

描述分布形状的两大基本属性是尺度和偏态（scale and skewness）。尺度一般通过标准差测量得到；标准差则通过数据值与均值的差的二次函数计算得到。对于对称分布而言，解释这种测量是很容易的，但当分布变得高度不对称时，对它的解释便站不住脚。同样的情况也出现在重尾分布上。由于许多用以描述社会现象的分布是偏态的或者是重尾的，使用标准差来表示它们的尺度便会出现问题。放弃使用标准差来捕捉分布的离散程度，我们可以使用选定 p 值下的分位差尺度测量法（quantile-based scale measure）（QSC）：

$$QSC^{(p)} = Q^{(1-p)} - Q^{(p)} \text{ for } p < 0.5 \qquad [2.1]$$

我们可以通过 $Q^{(0.025)}$ 和 $Q^{(0.975)}$ 获得总体中间位置的 95%分布，或者通过 $Q^{(0.25)}$ 和 $Q^{(0.75)}$ 获得中间位置的 50%分布（即传统意义上的四分位距），或者任何中间部分的 $100(1-2p)\%$ 分布。

QSC 不仅提供了直接有效的尺度测量方法，而且促进了基于模型的尺度变化的测量方法的发展（详见第 5 章）。相比之下，根据由标准差测量的尺度变化来析出预测变量效应的模型方法，限制了我们发现其他可能的模式。

测量分布形状的第二种属性是偏态。这一属性是许多不平等研究的核心所在。偏态通过数据值与均值的差的三次函数来计算。当数据围绕样本均值对称分布时，偏态的值等于 0。负的偏态值对应左偏分布；反之，正的偏态值则对应右偏分布。换言之，偏态的存在，表明中位数以下的分布和中位数以上的分布出现了不平衡。

尽管偏态常常被用作描述分布的非正态性，但是偏态需要通过分布的高阶矩计算得到，这一事实限制了偏态的作用。我们需要寻找连接属性（如偏态）和协变量的更灵活的方法。与动差法（moment-based measure）相比，样本分位数可以在许多方面描述分布的非正态性。分位数和分布形状的简单结合有助于进一步发展出模型化形状变化的方法（这一方法将在第 5 章讨论）。

非对称的上端和下端部分可通过分位数函数表达。图 2.6 分别描述了正态分布和右偏分布的分位数函数。正态分布的分位数函数围绕第 0.5 分位数（中位数）对称分布。例如，见图 2.6a，分位数函数在第 0.1 分位数的切线斜率等于在第 0.9 分位数的切线斜率。其他低-高对应的分位数同样具备这一特性。相比之下，偏态分布的分位数函数围绕中位数的分布则是不对称的。例如，见图 2.6b，分位数函数在第 0.1 分位数的切线斜率明显不同于在第 0.9 分位数上的切线斜率。

用上端部分（upper spread）表示中位数之上的分布情况，而下端部分（lower spread）对应于中位数之下的分布。对于对称分布而言，上端部分和下端部分是相同的。另一方面，在右偏分布中，下端部分远远短于上端部分。我们将分位差偏态测量（QSK）（quantile-based skewness）量化为上端部分与下端部分的比值减去 1。

$$QSK^{(p)} = (Q^{(1-p)} - Q^{(0.5)})/(Q^{(0.5)} - Q^{(p)}) - 1 \text{ for } p < 0.5 \qquad [2.2]$$

量化的 $QSK^{(p)}$ 通过减去 1 而重新回到中心位置，因此对于对称分布而言，它依然等于 0。当 $QSK^{(p)}$ 值大于 0 时，为右偏分布；当 $QSK^{(p)}$ 值小于 0 时，为左偏分布。

表 2.1 呈现了图 2.6 中的对称分布和右偏分布的 9 个分位数，上端部分和下端部分，4 个不同 p 值下的 $QSK^{(p)}$ 值。对于对称分布而言，$QSK^{(p)}$ 等于 0；而在右偏分布中，该值在 0.3 至 1.7 之间。$QSK^{(p)}$ 的定义是简单明了的，并且可扩展至测量由协变量导致的偏态变化（见第 5 章）。

到目前为止，我们根据 CDF 定义了分位数，并且发展出了分位差形状测量方法。另一种分位数的替代性定义有助于我们进一步理解分位数回归的估计

量(见下一章),我们建议对此感兴趣的读者继续阅读下一个部分,而其他读者可跳至小结部分。

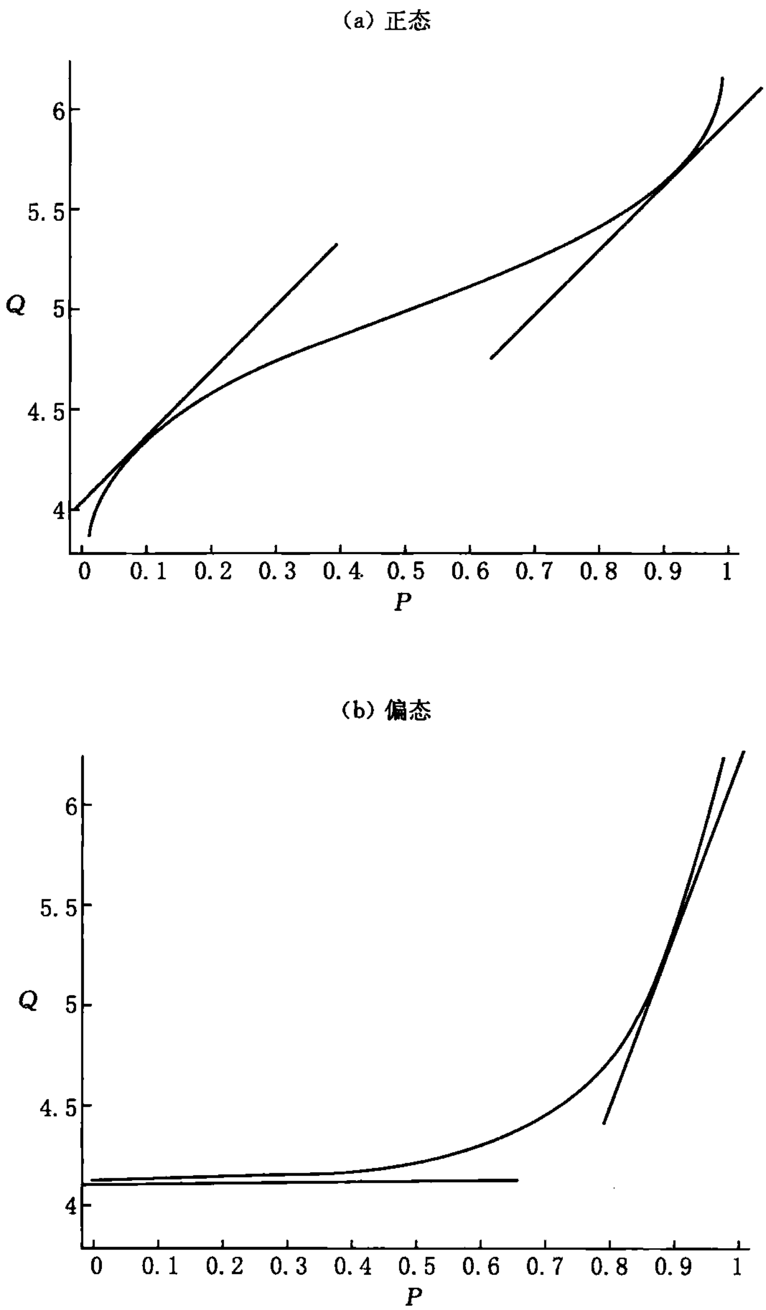


图 2.6 正态对偏态的分位数方程

表 2.1　分位差偏态测量方法

总体比例	对称分布			右偏分布		
	分位数	下端或者上端部分	QSK	分位数	下端或者上端部分	QSK
0.1	100	110	0	130	60	1.7
0.2	150	60	0	150	40	1.3
0.3	180	30	0	165	25	1.0
0.4	200	10	0	175	12	0.3
0.5	210	—	—	190	—	—
0.6	220	10	—	210	20	—
0.7	240	30	—	240	50	—
0.8	270	60	—	280	90	—
0.9	320	110	—	350	160	—

分位数作为某些最小化问题的解决方法

分位数同样可视为某些最小化问题的解决方法。由于下章将讨论分位数回归估计量的意义，所以我们需要引入新的定义。在这里，我们从中位数（第 0.5 分位数）开始讨论。

为介绍最小化问题，我们首先考虑大家熟悉的 y 分布的均值 μ。我们通过平方差 $(Y-\mu)^2$ 可以测量 Y 的某一数据点与均值 μ 的距离，然后通过期望平方差 $E[(Y-\mu)^2]$ 计算 Y 与均值 μ 的平均距离。

定义分布中心的一种方法是寻找 μ 值从而使得 Y 的均方差（average squared deviation）最小化。因此，我们有：

$$
\begin{aligned}
E[(Y-\mu)^2] &= E[Y^2]-2E[Y]\mu+\mu^2 \\
&= (\mu-E[Y])^2+(E[Y]^2-(E[Y])^2) \\
&= (\mu-E[Y])^2+\mathrm{Var}(Y) \qquad [2.3]
\end{aligned}
$$

因为第二项 Var(Y)是固定的，我们可通过最小化第一项 $(\mu-E[Y])^2$ 从而

最小化方程 2.3。当 $\mu = E[Y]$ 时，可使第一项等于 0，同时最小化方程 2.3，除此之外 μ 的其他值都将使第一项为正值而使方程 2.3 偏离最小值。

相似的，样本大小为 n 的样本均值同样可作为最小化问题的解决方法。我们寻找最小化平方均差 $\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2$ 的 μ 值：

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2 &= \frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{y})^2+\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\\ &= (\mu-\bar{y})^2+s_y^2 \end{aligned} \qquad [2.4]$$

这里，$\bar{y}$ 表示样本均值，s_y^2 表示样本方差。解决最小化问题就是使得第一项尽可能的小，那就是使 $\mu = \bar{y}$。

具体而言，我们考虑包括下列 9 个值的样本：0.23，0.87，1.36，1.49，1.89，2.69，3.10，3.82 和 5.25。样本数据点与特定 μ 值的均值平方差的描绘图如图 2.7a 所示。注意最小化函数的数值点位于平滑抛物线中间凸起的部分。

而中位数 m 也具有相似的最小化性质。相对于使用平方距离，我们可根据绝对距离 $|Y-m|$ 测量 Y 与 m 的距离，并且根据平均绝对距离(mean absolute distance) $E|Y-m|$ 计算总体中 Y 与 m 的平均距离。同样在最小化 $E|Y-m|$ 的情况下，我们得到值 m。我们知道，$|Y-m|$ 函数也是中部凸起的，因此，寻找某特定值，使得与 m 相关的导数等于 0 或者两个方向导数(directional derivatives)的正负符号不一致，这样 $|Y-m|$ 函数便实现了最小化。这个答案便是分布的中位数。(详细证明见本章的附录)

下面，我们讨论具体例子。我们将 m 至样本点的平均绝对距离定义为 $\frac{1}{n}\sum_{i=1}^{n}|y_i-m|$。对于上面提到的包括 9 个样本点的样本，它的平均绝对距离函数如图 2.7b 所示。对比于图 2.7a 所展示的函数图(均方差)，图 2.7b 的函数在外观上同样是中部凸起的抛物线。然而，图 2.7b 的函数并不是平滑的曲线，而是每个样本点上的切线斜率有明显变化的分段直线。图中函数的最小值与样本中位数 1.89 一致。这是众多现象中的特殊例子。对于任何样本，由

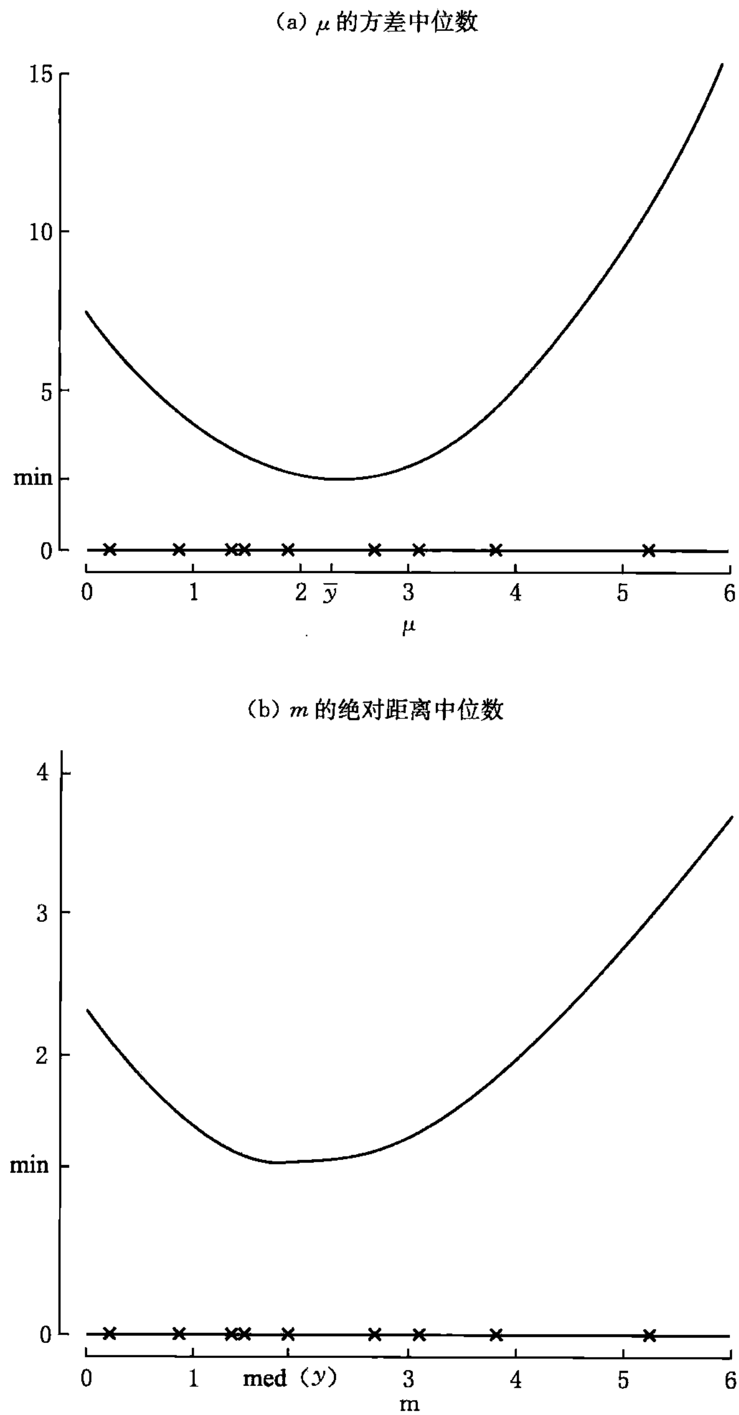


图 2.7 平均数和中位数作为最小化问题的解决方案

$f(m)=\frac{1}{n}\sum_{i=1}^{n}|y_i-m|$ 定义的函数是"V"形函数 $f_i(m)=|y_i-m|/n$ 的总和(见图 2.8 中,函数 f_i 对应数据点 $y_i=1.49$)。当 $m=y_i$ 时,函数 f_i 有最小值 0;并且当 $m<y_i$ 时,它的导数为 $-1/n$; 当 $m>y_i$ 时,其导数为 $1/n$。当该函数在 $m=y_i$ 上不可微时,在负向上有方向导数 $-1/n$, 在正向上有方向导数 $1/n$。作为这些函数的总和,f 对 m 在负向上的方向导数为 $(r-s)/n$, 在正向上则为 $(s-r)/n$; 这里 s 是 m 值右边的样本点数,r 是 m 值左边的样本点数。f 的最小值出现在当 m 的右边样本点数和左边样本点数相等时,即 m 为样本中位数。

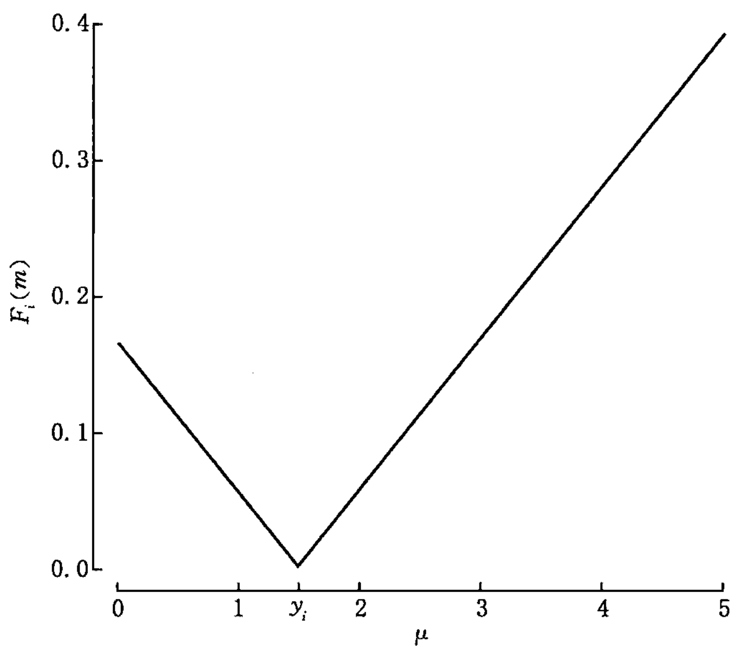


图 2.8 用于解决最小化问题的中位数的 V 形函数

对中位数的表述可推广至其他分位数上。对于任何 $p\in(0,1)$, Y 至特定 q 值的距离可由绝对距离测量,但根据 Y 是位于 q 的左边还是右边,我们赋予不同的权重。因此,我们将 Y 至特定 q 值的距离定义为:

$$d_q(Y,q)=\begin{cases}(1-p)\,|Y-q| & Y<q\\ p\,|Y-q| & Y\geqslant q\end{cases} \qquad [2.5]$$

我们寻找使 Y 的均值距离最小化的值 q:$E[d_p(Y, q)]$。当 q 是第 p 分位数时便达到最小值(见本章附录)。

同样的,第 p 样本分位数正是使平均(加权)距离最小化的 q 值。

$$\frac{1}{n}\sum_{i=1}^{n} d_p(y_i, q) = \frac{1-p}{n}\sum_{y_i<q} | y_i - q | + \frac{p}{n}\sum_{y_i>q} | y_i - q |$$

分位数的性质

分位数的一个基本性质是单调同变性(monotone equivariance)。如果我们对一个随机变量进行函数 h 的单调转换(如指数或对数函数),分位数可通过对分位数函数进行同样的转换而得利。换言之,如果 q 是 Y 的第 p 分位数,那么$h(q)$是 $h(Y)$的第 p 分位数。同样的情况存在于样本分位数中。例如,对于一个样本数据,由于我们知道第 20 百分位数是 4300,如果我们对数据进行对数转换,转换后的数据中第 20 百分位数就等于 log(4300) = 8.37。

样本分位数的另一个基本特性是他们对离群值的不敏感性。这一特性类似于分位数回归的特性,将使分位数和分位差程序(quantile-based procedures)在许多情况中变得实用。假如有中位数为 m 的样本数据 x_1, …, x_n,我们将一个位于中位数之上的数据值 x_i 替换成同样在中位数之上的其他值,从而修改了样本。同样的,我们也可以将一个位于中位数之下的数据值替换成同样在中位数之下的其他值。这样的修改对样本中位数[1]没有任何影响。任意第 p 样本分位数同样拥有相似的特性。

我们将这种情形与样本均值进行比较:将样本值 x_i 改为其他值 $x_i + \Delta$,将使样本均值变化 Δ/n。这样,个别数据点对样本分位数的影响是有限的,但对样本均值的影响则并非如此。

小结

本章介绍了分位数和分位数回归等概念。我们通过分布函数定义分位数

和分位数函数,并发展出测量分布位置和形状的分位差测量方法;还通过与传统分布矩(conventional distribution moments)相比较,强调了它们的实用性。另外,为了让读者更加了解分位数回归的估计量,我们将分位数重新定义为最小化问题的解决方法。有了以上准备,我们进入下一章讨论分位数回归模型及其估计量。

第 2 章 | 附录

证明:解决最小化问题的中位数和分位数

为了简述起见,我们假设分布函数 F 具有概率密度函数 f。为探讨中位数为什么可被定义为最小化问题,我们可以写成:

$$
\begin{aligned}
E\mid Y-m\mid &= \int_{-\infty}^{+\infty} \mid y-m\mid f(y)dy \\
&= \int_{y=-\infty}^{m} \mid y-m\mid f(y)dy + \int_{y=m}^{+\infty} \mid y-m\mid f(y)dy \\
&= \int_{y=-\infty}^{m} (m-y)f(y)dy + \int_{y=m}^{+\infty} (y-m)f(y)dy \qquad [A.1]
\end{aligned}
$$

如图 2.7b 所示,方程 A.1 是一个凸函数。实现最小化的解决方法是对 m 求微分并使这一偏导方程等于 0。第一项式的偏导数如下:

$$
\begin{aligned}
\frac{\partial}{\partial m}\int_{y=-\infty}^{m}(m-y)f(y)dy &= (m-y)f(y)\mid_{y=m} + \int_{y=-\infty}^{m}\frac{\partial}{\partial m}(m-y)f(y)dy \\
&= \int_{y=-\infty}^{m} f(y)dy = F(m)
\end{aligned}
$$

而第二项式的偏导数如下:

$$
\frac{\partial}{\partial m}\int_{y=m}^{+\infty}(y-m)f(y)dy = -\int_{y=m}^{+\infty} f(y)dy = -(1-F(m))
$$

合并以上两个偏导数,则有:

$$
\frac{\partial}{\partial m}\int_{-\infty}^{+\infty} \mid y-m\mid f(y)dy = F(m)-(1-F(m)) = 2F(m)-1 \qquad [A.2]
$$

通过设定 $2F(m)-1=0$,我们得到 $F(m)=1/2$,即是中位数,来满足最小

化要求。

重复上述关于分位数的计算，与方程 A. 2 相对应的分位数偏导数为：

$$\frac{\partial}{\partial q}E[d_p(Y, q)] = (1-p)F(q) - p(1-F(q)) = F(q) - p \qquad [A.3]$$

我们设定偏导数 $F(q) - p = 0$ 以得到 $F(q) = p$ 进而满足最小化要求。

第 3 章 | 分位数回归模型及其估计量

第 2 章讨论的分位数函数足以用来描述和比较单变量的分布情况。然而，当需要对一个因变量与一组自变量的关系进行建模时，我们必须在分位数函数的基础上引入回归，即分位数回归模型（QRM）。对于一组协变量，线性回归模型（LRM）代表条件均值函数，而 QRM 代表条件分位数函数。以 LRM 为参照标准，本章介绍 QRM 及其估计量，比较 LRM 和 QRM 的基本模型设置，LRM 的最小二乘估计和 QRM 的类似估计方法，还有两类模型的其他特性。下面，我们将通过分析家庭收入[2]的实证例子来阐明本章的基本观点。

线性回归模型及其短处

在社会科学研究中，LRM 是广泛使用的标准统计模型，但它只关注因变量的条件均值，而没有充分考虑因变量条件分布的完整特征。相比之下，QRM 有助于我们分析因变量条件分布的完整特征。在某些方面，QRM 和 LRM 是相似的，两种模型都可处理具有线性未知参数的连续型因变量，但 QRM 和 LRM 对不同的量进行建模，并且依赖于对误差项的不同假设。为了更好地理解这些相似点和不同点，我们首先讨论 LRM，然后再介绍 QRM。为了解释方便，我们集中讨论单一协变量的例子。当扩展至多个协变量时，虽然会增加额外的复杂性，但其思想在本质上是相同的。

假设 y 为依赖于 x 的一个连续型因变量。在我们的实证例子中，家庭收入是因变量。对于 x，我们用定距变量 *ED*（户主的受教育年限）表示，或者用虚拟变量 *BLACK*（户主的种族，黑人为 1，白人为 0）代替。在微观单位（在本例中是

家庭户)的样本中,数据是以 (x_i, y_i) 的形式成对出现的,其中 $i = 1, \cdots, n$。

通过 LRM,标准线性回归模型可表示为:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad [3.1]$$

在这里,ε_i 服从均值为 0、方差 σ^2 未知的正态独立同分布。ε_i 的均值假设为 0 的结果是,我们可看到拟合数据的函数 $\beta_0 + \beta_1 x$ 与 x 取特定值时的 y 的条件均值(表示为 $E[y|x]$)相对应,这可被理解为与协变量 x 特定值相对应的总体 y 值的平均数。

例如,当我们以受教育年限为协变量,进而拟合线性回归方程 3.1 时,我们得到预测方程 $\hat{y} = -23127 + 5633ED$,因此将选定的受教育年限代入方程中,将得到以下有关收入的条件均值。

ED	9	12	16
$E[y\|ED]$	\$ 27570	\$ 44469	\$ 67001

假设这是一个完美的拟合,我们便可以将这些数值理解为拥有特定受教育年限的个人所获得的平均收入。例如,受过 9 年教育的个人的平均收入为 27570 美元。

相似的,当协变量为 *BLACK* 时,这一拟合回归方程为 $\hat{y} = 53466 - 18268BLACK$,当代入协变量的数值时,我们得到以下数值。

BLACK	0	1
$E[y\|BLACK]$	\$ 53466	\$ 35198

假设上述拟合模型确实反映了总体真实情况,我们可将这些值当作子总体的平均值,例如白人家庭的平均收入为 53466 美元,而黑人家庭则为 35198 美元。

因此,我们得知了线性回归模型的一个基本原则,那就是它通过利用分布的均值来表示其集中趋势,从而尝试描述条件分布的位置的变化情况。LRM 的另一个基本特征是要求满足方差齐性假设;即假设条件方差 $\mathrm{Var}(y|x)$ 对于协变量的所有取值都等于常数 σ^2。当方差齐性假设不满足时,可以通过同时对

条件均值和条件尺度进行建模而调整 LRM。例如，我们可以调整方程 3.1 从而对条件尺度进行建模：$y_i = \beta_0 + \beta_1 x_i + e^{\gamma}\varepsilon_i$，这里 γ 是另一个未知的参数，我们可写成 $\mathrm{Var}(y \mid x) = \sigma^2 e^{\gamma}$。

因此，对 LRM 的运用，反映了协变量与因变量关系的某些特征，并可用于完成模型化尺度变化的任务，而尺度变化被认为是条件分布最重要的形状改变。然而，对条件尺度的估计并不总是可以通过统计软件轻易获得。另外，线性回归模型对建模者设有明显的限制，并且，使用 LRM 来模型化更加复杂的条件形状变化，是很具挑战性的。

为了描述采用 LRM 难以模型化的形状变化类型，各位可想象一些极端情况。对于大家感兴趣的一些总体，我们有一个因变量 y 和协变量 x，它们的特征是在 $x = 1, 2, 3$ 时，y 的条件分布拥有图 3.1 所示的概率密度。该图上的三种概率密度函数拥有相同的均值和标准差。由于因变量 y 的条件均值和尺度并不随着 x 而变化，所以通过对这些总体中的样本拟合一个线性回归模型，并不能提供有用的信息。为了理解协变量是如何影响因变量的，我们需要一个新的工具。分位数回归便是完成这一任务的恰当方法。

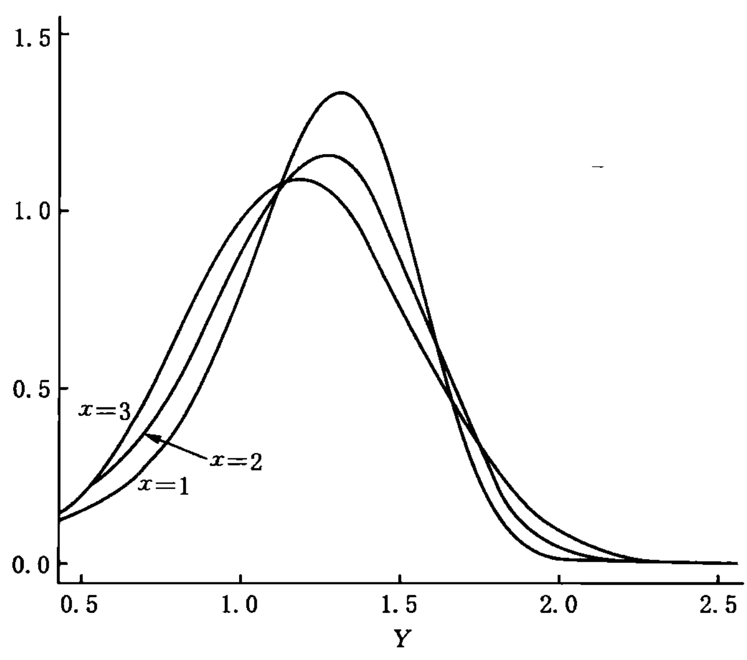


图 3.1 平均数和标准差相同但偏斜不同的条件分布

LRM 的第三个显著特征是它的正态性假设。因为 LRM 确保传统的最小二乘法可以最好地拟合数据，如果不作正态性假设，我们也可以通过 LRM 达到纯粹描述的目的。然而，在社会科学研究中，LRM 主要用来检验解释变量是否对因变量具有显著影响。假设检验需要超越参数估计，并且要求确定估计量的抽样偏差。p 值的计算依赖于正态性假设或者大样本近似值。如果违反这些条件，那么 p 值可能是有偏的，并且将导致无效的假设检验。

另一个与 LRM 相关的假设，是要求使用的回归模型适用于所有数据，我们称之为单一模型假设(one-model assumption)。LRM 中的离群值(不符合数据大多数情况的值)将对拟合的回归直线产生不良的影响。通常的做法是在 LRM 中确定并去除这些离群值。但离群值的意义及去除离群值的操作会破坏许多社会科学研究的结论，尤其是关于社会分层和不平等的研究，在这里，离群值和它们相对于多数数据值的位置是研究的重点。在建模方面，我们可能需要同时模型化大多数个案的关系和模型化离群值的关系，这一任务是 LRM 不可能完成的。

我们将用家庭收入数据举例说明以上提到的所有特性：以分布的观点看到的条件均值的局限，对方差齐性假设、正态性假设及单一模型假设的违反。图 3.2 展示了不同教育和种族群体的收入分布。收入的位置在三类不同教育群体之中和黑人与白人之间的变化是明显的，并且它们的形状也存在重要差异。因此，LRM 的条件均值无法表述由于协变量(教育或种族)的改变而导致的形状变化。另外，由于不同教育群体和两大种族的收入分布存在实质性差别，方差齐性假设便无法满足，而且也无法准确地估计标准误。在图 3.2 中的箱线图都是右偏的。条件均值和条件尺度模型不可能检测这些类型的形状变化。

通过检验残差图(residual plot)，我们可以识别 7 个离群值，包括受教育年限为 18 而收入超过 505215 美元的 3 个个案，还有教育年限为 20 而收入超过 471572 美元的 4 个个案。当在教育—收入模型中增加代表这些离群值阶层身份的虚拟变量时，我们发现这些个案使得截距额外增加了 483544 美元。

这些结果表明 LRM 方法的不足是由多方面的原因造成的，其中包括异方差性与离群值假设和对多种形状变化形式的检测失败。这些不足不仅出现在

关于家庭收入的研究中，并且当采用其他测量方法时，也会存在。因此，我们需要一个替代性的方法，用来处理异方差性和离群值，并检测形状变化的多种形式。

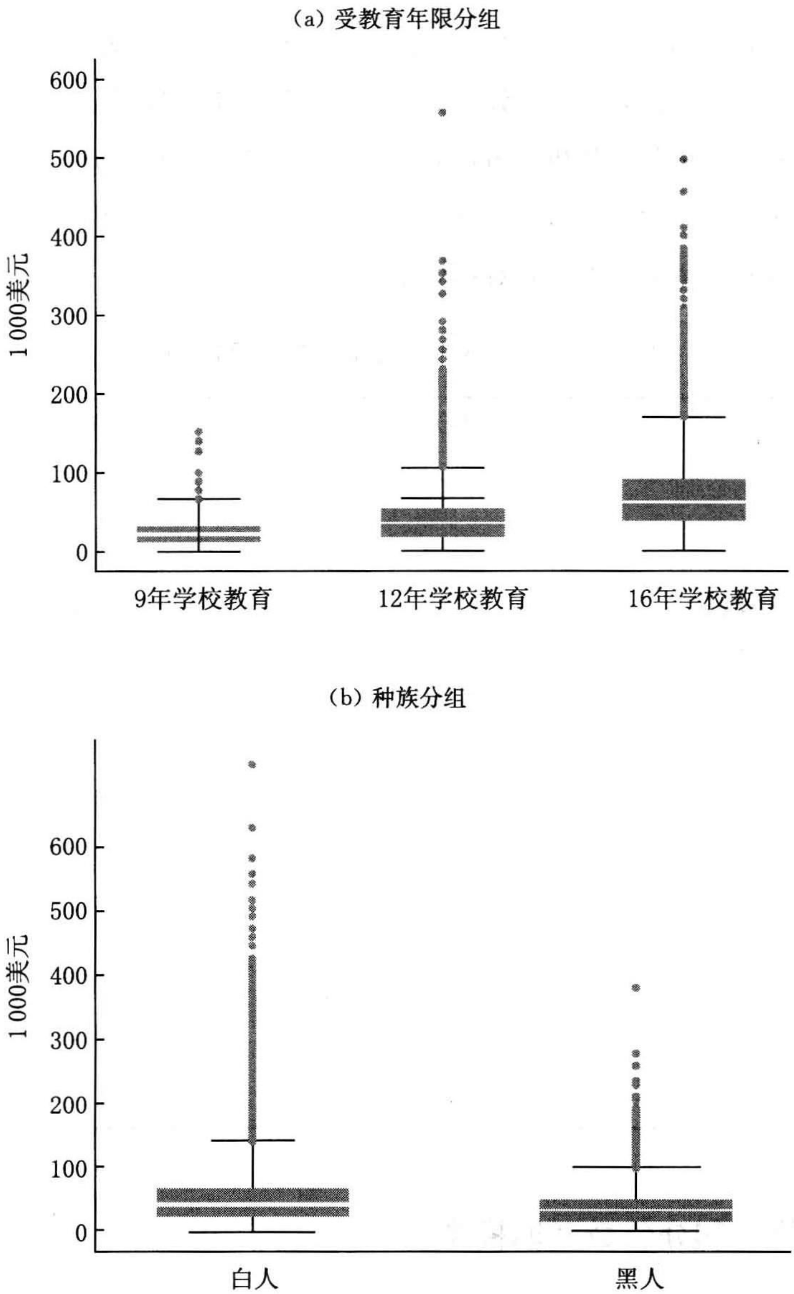


图 3.2 家庭收入的图示

正如上面提到的，条件均值无法识别形状变化。条件均值模型并不总是可以正确地模型化中心位置的变化，尤其是当因变量的分布不对称时。对于一个对称的分布，均值和中位数是一致的，但偏态分布的均值和中位数（第 0.5 分位数）则是不一致的。表 3.1 描述了家庭收入分布的一组简单统计数字。分布的右偏态性使得均值远大于中位数，不论是在整个样本中还是教育或种族的子样本中（见表 3.1 的首两行）。当分布的均值和中位数不一致时，中位数可以更加准确地表示分布的集中趋势。当我们检验中位数而不是均值时，三类不同教育群体和黑人与白人之间的收入位置的变化幅度会小很多。这一差异增加了我们使用条件均值方法模型化非对称分布的位置变化时的忧虑。

表 3.1 家庭收入分布(总体，教育组和种族组)

	总体	$ED=9$	$ED=12$	$ED=16$	*WHITE*	*BLACK*
均值	50334	27841	40233	71833	53466	35198
分位数						
中位数						
（第 0.5 分位数）	39165	22146	32803	60545	41997	26763
第 0.1 分位数	11022	8001	10510	21654	12486	6837
第 0.25 分位数	20940	12329	18730	36802	23198	13412
第 0.75 分位数	65793	36850	53075	90448	69680	47798
第 0.9 分位数	98313	54370	77506	130981	102981	73030
分位差尺度（QBC）						
$(Q_{0.75}-Q_{0.25})$	44853	24521	34344	53646	46482	34386
$(Q_{0.9}-Q_{0.1})$	87291	46369	66996	109327	90495	66193
分位差偏态（QBS）						
$\frac{(Q_{0.75}-Q_{0.5})}{(Q_{0.5}-Q_{0.25})}-1$	0.46	0.50	0.44	0.26	0.47	0.58
$\frac{(Q_{0.9}-Q_{0.5})}{(Q_{0.5}-Q_{0.1})}-1$	1.10	1.28	1.01	0.81	1.07	1.32

条件中位数和分位数回归模型

对于一个偏态分布，采用中位数测量集中趋势是更适合的，因此，条件中位

数回归(conditional median regression)而不是条件均值回归,应该被考虑用作模型化位置的变化。条件中位数回归模型在 18 世纪中期由 Boscovich 首次提出,随后由 Laplace 和 Edgeworth 加以研究。中位数回归模型弥补了 LRM 条件均值估计的不足。中位数回归估计的是协变量对条件中位数的影响效应,所以即使分布是偏态的,它也可以代表分布的中心位置。

为了模型化位置变化和形状变化,凯恩克和巴西特(1978)提出了比中位数回归模型更为一般化的模型——分位数回归模型(QRM)。QRM 估计了协变量潜在的微小变化对条件分布中各种不同的分位数的影响,例如,从第 0.05 分位数到第 0.95 分位数之间的 19 个等距分位数。通过中位数和非中位数的分位数,这 19 条拟合的回归线可以捕捉位置的变化(中位数的回归线),还有尺度和更复杂的形状变化(非中位数的回归线)。因此,QRM 可以估计协变量对整个分布的微小影响,并且顾及到了异方差性。

根据凯恩克和巴西特的设定,与方程 3.1 中的 LRM 相对应的 QRM 可表示为:

$$y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i + \varepsilon_i^{(p)} \qquad [3.2]$$

在这里 $0<p<1$ 表示数值小于第 p 分位数的比例。回到 LRM,特定 x_i 值下的 y_i 的条件均值为 $E[y_i \mid x_i] = \beta_0 + \beta_1 x_i$,并且误差项 ε_i 的期望值为 0。而与之对应的 QRM,在特定 x_i 值下的第 p 条件分位数为 $Q^{(p)}(y_i \mid x_i) = \beta_0^{(p)} + \beta_1^{(p)} x_i$。这样,第 p 条件分位数是由特定分位数下的参数 $\beta_0^{(p)} + \beta_1^{(p)}$ 和协变量 x_i 的特定值决定的。正如 LRM,在关于误差项 ε_i 的假定下,QRM 同样可被公式化。由于 $\beta_0^{(p)} + \beta_1^{(p)} x_i$ 是固定的,如果要使 $Q^{(p)}(y_i \mid x_i) = \beta_0^{(p)} + \beta_1^{(p)} x_i + Q^{(p)}(\varepsilon_i) = \beta_0^{(p)} + \beta_1^{(p)} x_i$,则要求误差项的第 p 分位数等于 0。

需要重点留意的是,对于分位数 p 的不同值而言,特定 i 下的误差项 $\varepsilon_i^{(p)}$ 是彼此相关的。事实上,用 q 代替 p 后,方程 3.2 变成 $y_i = \beta_0^{(q)} + \beta_1^{(q)} x_i + \varepsilon_i^{(q)}$,从而有 $\varepsilon_i^{(p)} - \varepsilon_i^{(q)} = (\beta_0^{(q)} - \beta_0^{(p)}) + x_i(\beta_1^{(q)} - \beta_1^{(p)})$,所以这两个误差项在特定 x_i 下相差一个常数。换言之,$\varepsilon_i^{(p)}$ 的分布和 $\varepsilon_i^{(q)}$ 的分布只是彼此的一种位移(分布形状不变)。另一个需要考虑的关于 QRM 的重要的特殊情况是,当

$i=1, \cdots, n$ 时，$\varepsilon_i^{(p)}$ 服从独立同分布；我们将此视为 i. i. d 情况。在这一情形下，$\varepsilon_i^{(p)}$ 的第 q 分位数是依赖于 p 和 q 而不是 i 的常数 $C_{p,q}$。利用方程 3.2，我们可将第 q 条件分位数方程写成 $Q^{(q)}(y_i \mid x_i) = Q^{(p)}(y_i \mid x_i) + C_{p,q}$。[3]所以我们的结论是，在 i. i. d 情况下，当斜率 $\beta_1^{(p)}$ 换成一般值 β_1 时，条件分位数方程只是彼此的一种简单的位移情况。换言之，*i. i. d* 假设指的是因变量的分布不存在形状变化。

方程 3.1 中的 LRM 只有一个由一条方程式表示的条件均值，与之不同的是，方程 3.2 表明 QRM 拥有多个条件分位数。这样，多个方程可以方程 3.2[4]的形式表示。例如，如果 QRM 指定 19 个分位数，这 19 个方程会产生关于 x_i 的 19 组系数，分别表示 19 个条件分位数 ($\beta_1^{(0.05)}$, $\beta_1^{(0.10)}$, …, $\beta_1^{(0.95)}$)。这些分位数并不要求一定是等距的，但在实践中，将它们设为等距的会便于解释。

在我们的例子中，拟合方程 3.2 将产生在已知教育或种族情况下的收入的 19 个条件分位数的估计值(见表 3.2 和表 3.3)。教育的系数从第 0.05 分位数的 \$1019 单调递增至第 0.95 分位数的 \$8385。相似的，在低端分位数中的黑人种族效应小于在高端分位数的情况。

12 年受教育年限的条件分位数是：

p	0.05	0.50	0.95
$E(y_i \mid ED_i = 12)$	\$7976	\$36727	\$111268

黑人种族的条件分位数是：

p	0.05	0.50	0.95
$E(y_i \mid BLACK_i = 1)$	\$5432	\$26764	\$91761

这些结果与 LRM 条件均值的结果十分不同。条件分位数描述了一种条件分布，被用以概括位置和形状的变化。对 QRM 估计值的解释是第 5 章和第 6 章的话题。

表 3.2 教育-家庭收入对教育的分位数回归估计值

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ED	1 109	1 617	2 023	2 434	2 750	3 107	3 397	3 657	3 948	4 208
	(28)	(31)	(40)	(39)	(44)	(51)	(57)	(64)	(66)	(72)
常数	−4 252	−7 648	−9 170	−11 160	−12 056	−13 308	−13 783	−13 726	−14 026	−13 769
	(380)	(424)	(547)	(527)	(593)	(693)	(764)	(866)	(884)	(969)
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	
ED	4 418	4 676	4 905	5 214	5 557	5 870	6 373	6 885	8 385	
	(81)	(92)	(88)	(102)	(127)	(138)	(195)	(274)	(463)	
常数	−12 546	−11 557	−9 914	−8 760	−7 371	−4 227	−1 748	4 755	10 648	
	(1 084)	(1 226)	(1 169)	(1 358)	(1 690)	(1 828)	(2 582)	(3 619)	(6 101)	

注:括号内为标准误。

表 3.3 种族-家庭收入对种族的分位数回归估计值

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
BLACK	−3 124	−5 649	−7 376	−8 848	−9 767	−11 232	−12 344	−13 349	−14 655	−15 233
	(304)	(306)	(421)	(485)	(584)	(536)	(609)	(708)	(781)	(765)
常数	8 556	12 486	16 088	19 718	23 198	26 832	30 354	34 024	38 047	41 997
	(115)	(116)	(159)	(183)	(220)	(202)	(230)	(268)	(295)	(289)
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	
BLACK	−16 459	−17 417	−19 053	−20 314	−21 879	−22 914	−26 063	−29 951	−40 639	
	(847)	(887)	(1 050)	(1 038)	(1 191)	(1 221)	(1 435)	(1 993)	(3 573)	
常数	46 635	51 515	56 613	62 738	69 680	77 870	87 996	102 981	132 400	
	(320)	(335)	(397)	(392)	(450)	(461)	(542)	(753)	(1 350)	

注:括号内为标准误。

图 3.3 描述了从总样本抽取的大小为 1000 户家庭的随机样本情况，左图描述了基于户主受教育年限的家庭收入的散点图和 LRM 的拟合直线。这单一的回归直线描述了均值的变化，例如，从 12 年受教育年限的均值 $ 22532 转换为 16 年受教育年限的均值[5633×(16－12)]。但是，这条回归直线无法捕捉形状的变化。

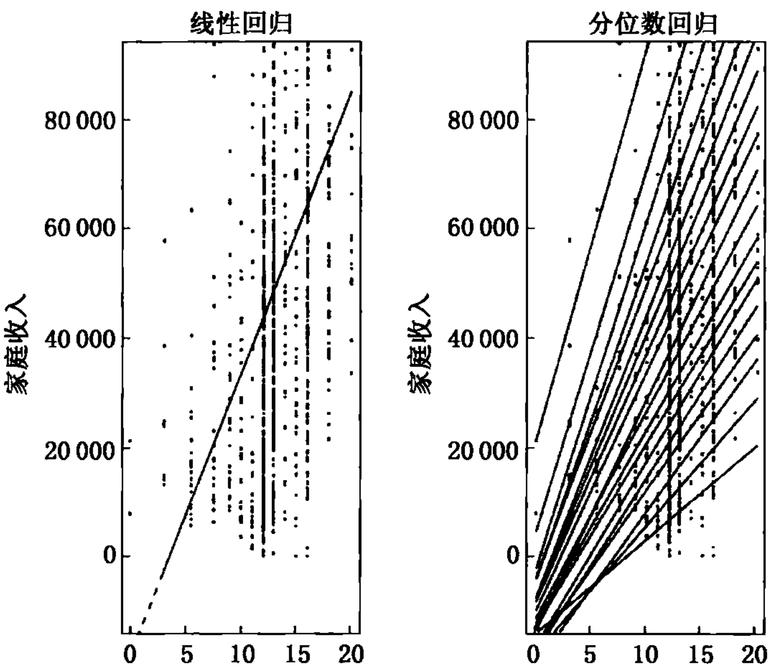


图 3.3 教育对家庭收入的条件平均数和条件分位数的影响(1000 个随机家庭样本)

图 3.3 中的右图呈现了与左图一样的散点分布和 19 条分位数回归线。50%分位数(中位数)的拟合线描述中心位置的变化，说明教育和条件中位数收入的正向关系。这一回归线的斜率为 $ 4208。当受教育年限从 12 增加至 16 时，收入增加的幅度达 $ 16832[4208×(16－12)]。这一变化小于 LRM 中的均值变化。

除了位置变化的估计外，其他 18 条分位数回归线提供了关于形状变化的信息。这些回归线都是正向的，但有着不同的斜率。这些回归线在低教育水平处密集(例如 0—5 年的受教育年限)，但在高教育水平处(如 16—20 年的受教育年限)，彼此的偏差越来越宽。通过低教育水平处的斜线密集和高教育水

平处的斜线分散情况，可以描述形状的变化。例如，受教育年限为 16 的条件收入（第 0.05 条件分位数上的 $ 12052 变为第 0.95 条件分位数上的 $ 144808）的变化幅度大于受教育年限为 12 的变化幅度（第 0.05 条件分位数上的 $ 7976 变为第 0.95 条件分位数上的 $ 111268）。因此，非中位数条件分位数将位置变化从形状变化中分离出来。这一特征在确定协变量对因变量条件分布的位置变化和形状变化的影响效应上是至关重要的。这一话题将在第 5 章和对 QRM 结果的解释一起讨论。

分位数回归（QR）估计

我们通过回顾最小二乘估计而将 QR 估计放在熟悉的情境中讨论。最小二乘估计通过计算最小化残差的参数值，从而实现 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的参数估计：

$$\min \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2 \qquad [3.3]$$

如果 LRM 假设是正确的，当样本容量趋向无穷时，这一拟合的响应函数 $\hat{\beta}_0 + \hat{\beta}_1$ 将接近总体的条件均值 $E(y \mid x)$。在方程 3.3 中，最小化的部分是数据点 (x_i, y_i) 到拟合直线 $y = \hat{\beta}_0 + \hat{\beta}_1 x$ 的垂直距离的平方总和。

一个类似的最小化方法可通过如下步骤实现：(a)分别对方程 3.3 在 β_0 和 β_1 处求偏导数；(b)设定每个偏导数都等于 0；(c)求解这个二元方程组。这样我们得到两个估计量：

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

与 LR 估计量明显不同的 QR 估计量的特点在于，在 QR 中，数据点到回归线距离的测量通过垂直距离的加权总和（没有平方）而求得，这里赋予拟合线之下的数据点的权重是 $1 - p$，而赋予拟合线之上的数据点的权重则是 p。对于 p 的每一个选择，如 $p = 0.10, 0.25, 0.50$，都会产生各自不同的条件分位数的

拟合函数。这一任务是为每一个可能的寻找适合的估计量。读者需要留意第 2 章中讨论过的，分布的均值可被看作最小化总体平均平方距离的值，而分位数 q 可被看作最小化平均加权距离的值——根据数据点在 q 值之上还是之下而进行加权。

具体而言，我们首先考虑中位数回归的估计量。在第 2 章中，我们描述了 y 的中位数 m 是如何最小化 $E\mid y-m\mid$ 的。与中位数回归例子相似的对策是，我们选择最小化绝对残差总和(the sum of absolute residuals)。换言之，我们需要寻找最小化绝对残差总和(观察值与拟合值之间的绝对距离)的系数。最小化方程 3.4 可得到估计量 βs：

$$\sum_i \mid y_i - \beta_0 - \beta_1 x_i \mid \qquad [3.4]$$

在适当的模型假设下，当样本大小趋于无限时，我们得到总体的特定 x 值下的 y 的条件中位数。

当方程 3.4 实现最小化时，我们得到一条中位数回归线(median-regression line)，这一直线穿越一对数据点，使得剩下数据的一半位于回归线上方，另一半则在下方。也就是：大概一半的残差是正的，另一半为负的。但是存在多条具有这种性质的回归线，只有通过最小化方程 3.4 才可得到中位数回归线。

算法细节

在这一部分，我们将讲解方程 3.4 的结构如何帮助我们寻找实现最小化的算法。对此不感兴趣的读者可跳过本部分。

图 3.4 中的左图展示了 8 组假设的数据 (x_i, y_i) 和 28 条 $\{[8(8-1)/2]=28\}$ 连结每一对数据点的直线。其中的虚线是拟合的中位数回归直线，即通过最小化所有数据的绝对垂直距离总和而得到的直线。由此观察到，存在 6 对数据点没有落在中位数回归线上，一半数据点在回归线的下方而另一半则位于其上方。经过平面(x, y)的每一个直线可表示截距—斜率为(β_0, β_1)的形式 $y=\beta_0+\beta_1 x$，因此在平面(x, y)上的直线和在平面(β_0, β_1)的点是相对应的。图 3.4

中的右图是(β_0，β_1)的平面图，图中包括左图每条直线所对应的点。特别的，中间的实心圆代表左图中的中位数回归线。

此外，如果截距和斜率为(β_0，β_1)的直线经过一个特定的点(x_i，y_i)，那么 $y_i = \beta_0 + \beta_1 x_i$，因此点($\beta_0$，$\beta_1$)落在直线 $\beta_1 = (y_i/x_i) - (1/x_i)\beta_0$ 上。这样，我们在平面(x，y)上的点和在平面(β_0，β_1)的直线之间建立了对应关系，反之亦然，这一现象被称为点-线二元性(point/line duality)(Edgeworth，1888)。

图 3.4 中的右图上的 8 条直线对应于左图上的 8 个数据点。这些直线将平面(β_0，β_1)分割为一个多边形区域。这种区域的例子如图 3.4 阴影部分所示。在任何一个区域中，那些数据点对应于平面(x，y)上的直线族，所有这些直线都以同样的方式将数据分为两个部分(即在一条直线上的数据点等同于另一条直线上的数据点)。因此，在方程 3.4 中我们设法最小化的(β_0，β_1)函数在每一个区域中都是线性的，所以这一函数随着一个形成多面体曲面(polyhedral surface)的图像而凸起，正如图 3.5 从两个不同角度为我们的例子所作的展示一样。

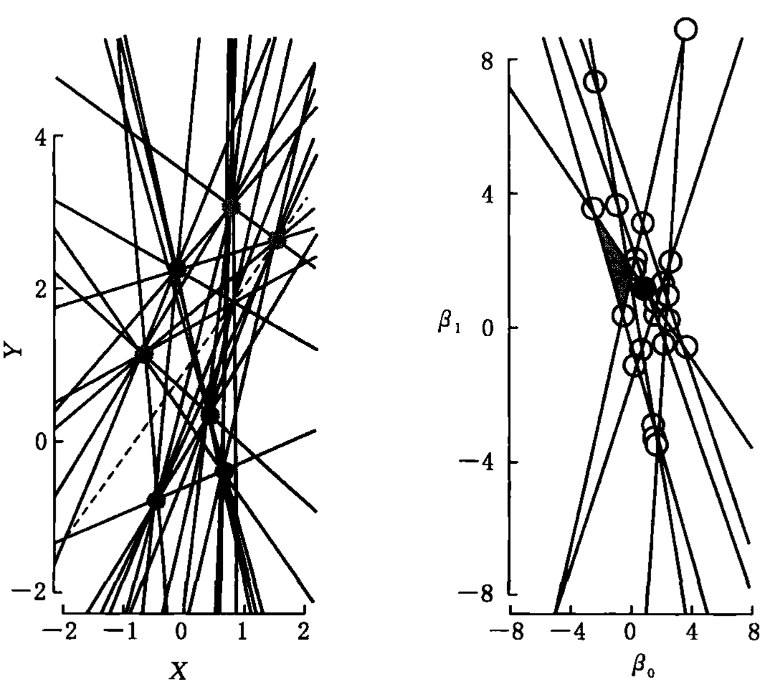


图 3.4 点线两重性的图示

这一多面体曲面的顶点、边和面分别投射出点、线段和区域，如图 3.4 中的右图展示的(β_0，β_1)平面。利用点线二元性的对应性，每一个顶点对应于连结数据对的一条直线。曲面上连结两个顶点的边则对应于一对这样的直线，在这里，定义第一条直线的其中一个数据点被另一个数据点代替，并且其他的数据点保持与两条直线的相对位置不变(上面或者下面)。

最小化方程 3.4 中绝对距离总和的计算方法，这种获得中位数回归系数($\hat{\beta}_0$，$\hat{\beta}_1$)的方法，可以建立在解决线性规划的外点算法(exterior-point algorithms)上。从对应于一个顶点的任意点(β_0，β_1)开始，沿着多面体曲面的边从顶点到顶点反复移动，最后选择在每个顶点处最倾斜的那条路径，直到实现最小化。

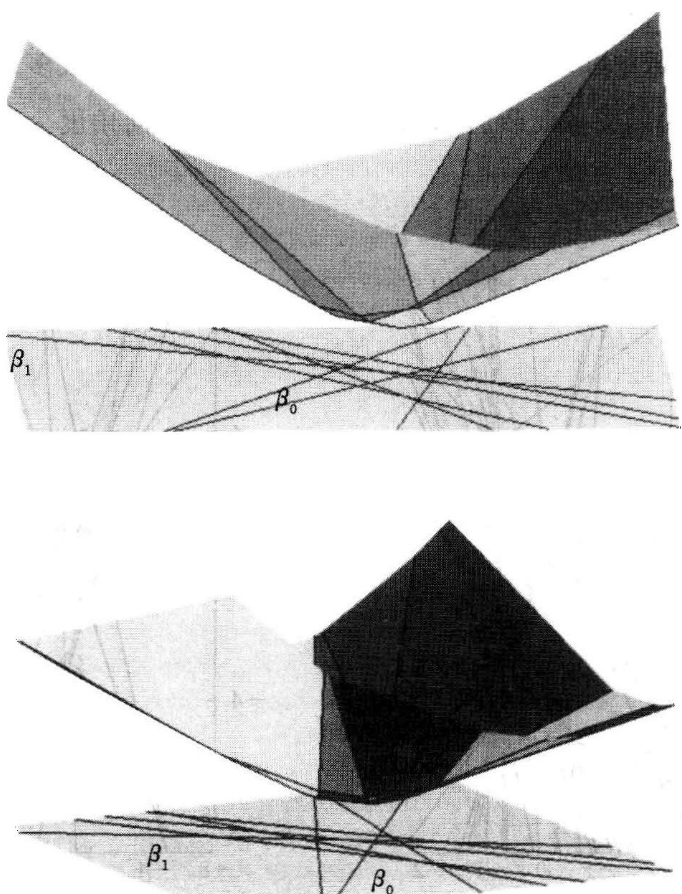


图 3.5 多面体表面及其投射

利用在之前段落中描述过的对应性特征，我们反复从由数据对确定的一条直线移动到另一条直线，在每一个步骤中确定当前的两个数据点到底是哪一个使得方程 3.4 达到最小值，这个数据点就被挑选出来。绝对误差总和的最小值是在低于曲面最低点的(β_0，β_1)平面上的点获得的。一个涉及在点 β_0 上的方向导数(类似于第 2 章提到的，中位数是解决最小化问题的数值)的小小争议可得出一个结论，即中位数回归线之上的数据点的数目等于回归线之下的数目。

中位数回归估计量可一般化为第 p 分位数回归的估计量(Koenker & d'Orey，1987)。回想第 2 章中的讨论，单变量样本 y_1，…，y_n 分布的第 p 分位数就等于使得样本数据中加权距离的总和最小化的 q 值，这里在 q 值之下的数据点的权重为 $1-p$，而在 q 值之上的数据点的权重为 p。在类似的情况下，我们将第 p 分位数回归估计量$\hat{\beta}_0^{(p)}$和 $\hat{\beta}_1^{(p)}$定义为可最小化拟合值 $\hat{y}_i=\hat{\beta}_0^{(p)}+\hat{\beta}_1^{(p)}x_i$ 与值 y_i 之间的加权距离总和的数值，这里，如果拟合值低估了观察值 y_i，则使用权重 $1-p$，在其他情况下则用权重 p。换言之，我们设法最小化残差 $y_i-\hat{y}_i$ 的总和，正向残差的权重为 p 而负向残差的权重为 $1-p$。正式的，第 p 分位数回归估计量$\hat{\beta}_0^{(p)}$和$\hat{\beta}_1^{(p)}$被用作最小化：

$$\sum_{i=1}^{n} d_p(y_i,\ \hat{y}_1) = p\sum_{y_i\geqslant\beta_0^{(p)}+\beta_1^{(p)}x_i} |\, y_i-\beta_0^{(p)}-\beta_1^{(p)}x_i\,|$$

$$+(1-p)\sum_{y_i<\beta_0^{(p)}+\beta_1^{(p)}x_i} |\, y_i-\beta_0^{(p)}-\beta_1^{(p)}x_i\,| \qquad [3.5]$$

这里 d_p 是第 2 章中介绍过的距离。因此，这一方程不同于方程 3.4，在方程 3.4 中正残差和负残差被赋予同等的重要性，而方程 3.5 给正残差和负残差赋予不同的权重。从方程 3.5 可观察，第一项的值等于位于回归线 $y_i=\beta_0^{(p)}+\beta_1^{(p)}x_i$ 上方的数据点到该线的垂直距离的总和；同样的，第二项则是回归线下方所有数据点的距离总和。

请注意：与常见的误解不同，每一个分位数回归的系数估计量是基于整体样本的加权数据计算的，而非基于这一分位数的部分样本。计算分位数回归系数 $\hat{\beta}_0^{(p)}$ 和 $\hat{\beta}_1^{(p)}$ 的方法，可依据类似勾勒中位数回归系数的直线而发展出来。第 p 分位数估计量拥有和中位数回归估计量相似的特性：位于拟合线 $y_i=\hat{\beta}_0^{(p)}+$

$\hat{\beta}_1^{(p)} x_i$ 之下的数据比例为 p，而位于上方的比例则为 $1-p$。

例如，当我们估计第 0.1 分位数回归线的系数时，位于回归线下方的观察值的权重为 0.9，而上方的观察值的权重为 0.1。结果，位于拟合线上方的 90%数据点(x_i，y_i)导致了正残差，而拟合线下方的 10%数据点拥有负残差。相反的，为了估计第 0.9 分位数回归的系数，赋予回归线下方的数据 0.1 的权重，其余则为 0.9 的权重；因此，90%观察值拥有负残差而剩下的 10%观察值拥有正残差。

转化与同变性

当分析一个因变量时，研究者常常为了解释的方便或者获得更好的模型拟合而变换尺度。模型和估计值的同变性表明，如果数据被转化了，模型和估计值也会经历同样的转化。当我们转化因变量后，同变性的知识有助于我们重新解释拟合模型。

对于因变量的任何线性转化，即在 y 的基础上加上一个常数或者乘以一个常数，LRM 的条件均值同样得到精确转化。这一陈述的正确性是基于以下事实：对于任意常数 a 和 c，我们有：

$$E(c+ay \mid x) = c + aE(y \mid x) \qquad [3.6]$$

例如，如果总体中每户家庭从政府那里获得 500 美元，总体的条件均值将同样增加 500 美元(新的截距将增加 500)。如果收入的单位从 1 美元转化为 1000 美元，单位为 1 美元的条件均值同样将增加到 1000 倍(截距和斜率将同时乘以 1000 以改变美元的单位)。相似的，如果工资价格已经由美元单位转化为美分单位，那么将条件均值(截距和斜率)除以 100 将再次转化为美元单位。这一性质称为线性同变性(linear equivariance)，因为线性转化对于因变量和条件均值而言是相同的。QRM 具有同样的性质：

$$Q^{(p)}(c+ay \mid x) = c + a(Q^{(p)}[y \mid x]) \qquad [3.7]$$

这里我们假设 a 是一个正向的常数。如果 a 为负，因为顺序相反，所以我们有：

$$Q^{(p)}(c+ay \mid x)=c+a(Q^{(1-p)}[y \mid x])$$

非线性转化是常常出现的情况。对数转化经常被用于处理分布的右偏态问题。为了使分布更加正态化或者获得更好的模型拟合，其他转化方法也会被考虑。

为了模型化协变量的相对效应（如百分比变化），对数转化同样被采用。换言之，协变量的效应更多地被看作是倍增的而不是递增的。在我们的例子中，教育和种族的影响效应在之前表达为递增的形式（美元单位），并同样可以用倍增的形式测量，如以百分比变化的形式。例如，我们可以提问：当受教育年限增加一年时，条件均值收入的百分比是如何变化的？当受教育年限增加一年时，在对数收入方程（乘以 100）中的教育系数近似于条件均值收入的百分比变化。然而，在 LRM 中，对数收入的条件均值并不等同于条件均值收入的对数。对收入和对数收入的 LRM 进行估计，产生以下两个拟合模型：

$$\hat{y}=-23127+5633ED,\ \log\hat{y}=8.982+0.115ED$$

从对数收入模型的结果可知，当教育增加一年时，条件均值收入将增加 11.5%。[5]而受教育年限为 10 年时，收入模型中的条件均值为 $33203，其对数为 8.108。在同等教育水平下，对数收入模型中的条件均值则为 10.062，这一数字大于收入条件均值的对数（8.108）。LRM 的因变量的对数转化允许将 LRM 的估计值解释成百分比变化，但是绝对形式的因变量条件均值不可能从对数化的条件均值中获得：

$$E(\log y \mid x) \neq \log[E(y \mid x)] \text{ 和 } E[y_i \mid x_i] \neq e^{E[\log y_i \mid x_i]} \qquad [3.8]$$

特别的，如果我们的目标是为了估计教育的绝对效应，我们需要选择收入模型；但如果是为了估计教育的相对效应，我们则选择对数收入模型。尽管这两个目标彼此关联，但是在任何简单的转化后，两种模型的条件均值并不相关。[6]因此，采用对数收入的结果对收入的分布作出结论将会犯错误（尽管这在实践中普遍存在）。

对数转化是单调转化方法族中的一种，那就是维持顺序不变的转化。正式

的，在 $y<y'$ 时，如果 $h(y)<h(y')$，那么转化形式 h 就是单调的。对于数值恒为正的变量而言，在常数 ϕ 为固定的正值时，幂转化 $h(y)=y^{\phi}$ 也是单调的。由于非线性影响，当我们对因变量进行单调转化后，转化后因变量的变化幅度对于不同的 y 值是不一样的。然而方程 3.6 的特性只在线性函数下才正确，对于一般的单调函数则不适用，那就是，$E(h(y)\mid x)\neq h(E(y_i\mid x_i))$。一般而言，"单调同变性"的性质对于条件均值来说是不成立的，所以 LRM 并不具有单调同变性。

相比之下，条件分位数具有单调同变性的特征；那就是对于一个单调函数 h，我们有：

$$Q^{(p)}(h(y)\mid x)=h(Q^{(p)}[y\mid x]) \tag{3.9}$$

这一性质直接遵循着第 2 章中提到的单变量分位数的单调同变性特征。尤其是，对数 y 的条件分位数就是 y 的条件分位数的对数：

$$Q^{(p)}(\log(y)\mid x)=\log(Q^{(p)}[y\mid x]) \tag{3.10}$$

同样的，

$$Q^{(p)}(y\mid x)=e^{Q^{(p)}[\log(y)\mid x]} \tag{3.11}$$

因此，我们可以将非转化变量的拟合分位数回归模型重新解释为转化变量的分位数回归模型。换言之，假设第 p 分位数函数的完美拟合形式为 $Q^{(p)}(y\mid x)=\beta_0+\beta_1x$，所以我们就有 $Q^{(p)}(\log y\mid x)=\log(\beta_0+\beta_1x)$，因此，我们可以用协变量的绝对效应描述协变量的相对效应，反之亦然。

以条件中位数作为例子：

$$Q^{(0.50)}(y_i\mid ED_i)=-13769+4208ED_i,\ Q^{(0.50)}(\log(y_i)\mid ED_i)$$
$$=8.966+0.123ED_i$$

受教育年限为 10 年时的收入条件中位数为 28311 美元。这一条件中位数的对数为 10.251，近似于在同等教育水平下对数收入方程的条件中位数为 10.196。相应的，当对数形式转变回初始形式，对数收入方程在受教育年限为 10 时，其条件中位数为 $e^{10.196}=28481$。

QRM 的单调同变性对于涉及偏态分布的研究而言是至关重要的。如果采用 LRM,原先的分布形状会在对数单位估计值反向转化时被扭曲,但是如果使用 QRM,就可以维持原先的分布。不平等研究中经常使用百分比变化的形式来表达协变量对因变量的影响。因此,单调同变性可以帮助研究者实现两个目标:测量协变量的单位变化而引起的因变量的百分比变化、测量这一变化对初始单位下的条件分布的位置和形状的影响。

稳健性

稳健性指的是:关于数据 y,对离群值的存在和模型假设的违反不敏感的一种性质。离群值被定义为违反 y 中多数值之间关系的一些数值。在 LRM 中,估计值受离群值影响较大。在前文中,我们举例说明了收入分布的离群值是如何扭曲均值和条件均值的。LRM 对离群值的高度敏感性是众所周知的。然而,在实践中,删除离群值不能满足许多社会科学研究,尤其是不平等研究的需要。

相比之下,QRM 估计值则不受离群值的影响。[7]它的稳健性由于最小化方程 3.5 中距离函数的性质而得到了增强,并且我们能够说明分位数回归估计值的特性,这类似于第二章关于单变量分位数的陈述。如果我们修改位于拟合分位数回归线之上(或之下)的某个数据点所对应的因变量的值,只要这个数据点依然处于回归线之上(或之下),这条拟合回归线将保持不变。换言之,如果我们在不改变残差的正负符号情况下修改因变量的值,这一拟合线将保持不变。在这种情况下,正如单变量分位数一样,离群值的影响是十分有限的。

另外,由于估计值的协方差矩阵是在正态假设下计算得到的,所以 LRM 的正态假设对于获取 LRM 的推论统计值是必需的。对于正态假设的违反将产生不准确的标准误。QRM 对分布假设的稳健性,是因为它的估计量更多地依赖于特定分位数附近的分布状态,而非远离分位数的分布情况。QRM 的推论统计值是不受分布情况影响的(这是本书第 4 章讨论的话题)。稳健性

在研究高度偏态的分布现象，如收入、财富、教育和健康结果时，是十分重要的。

小结

本章介绍了分位数回归模型的基本原理，并与线性回归模型进行了比较，包括模型设置、估计方法和估计值的性质等。QRM 延续了第二章介绍的样本分位数的许多特性。另外，我们还解释了 LRM 在特定情况下如何不适用于反映协变量对因变量分布的作用。同时，我们强调了 QRM 的一些关键特征，并展示了 QRM 与 LRM 之间众多重要的差异：(a)多元分位数回归对数据的拟合与一元线性回归分别对数据的拟合；(b)最小化残差绝对值的加权总和的分位数估计与最小化平方总和的最小二乘估计；(c)条件分位数对于分布假设的单调同变性和稳健性，而条件均值则缺乏这些特性。通过这些基本原理，我们可以进入关于 QRM 推论的话题。

第 4 章 | 分位数回归的推论

第 3 章涉及参数估计的话题。我们现在转向推论统计部分，尤其是 QRM 的标准误和系数估计值的置信区间。首先，我们综述 LRM 的推论方法，讨论有限样本情况和用于构建置信区间与检验假设的量的渐近分布情况。然后，我们介绍相应的 QRM 的渐近程序(asymptotic procedure)。接着，我们介绍 QRM 的自举程序——用于进行 QRM 系数的推论。因为关于渐近程序的假设常常不能成立，所以自举程序是更合适的；而且，即使这些假设得到满足，计算构造尺度和偏态变化的标准误也将十分复杂。自举程序为获取任意估计值或估计值组合的标准误和置信区间提供了灵活性。本章的最后部分将讨论关于拟合优度和模型诊断的话题。

LRM 的标准误和置信区间

首先，我们综合回顾一下 LRM 系数的推论过程，在理想的建模假设下，其表达式为 $y_i = \sum_{j=1}^{k}\beta_j x_j^{(i)} + \varepsilon_i$，这里的误差 ε_i 服从均值为 0、方差为常数 σ^2 的正态独立同分布(i. i. d)，因此这种分布是可导的。表达式 $x_j^{(i)}$ 表示为第 j 个协变量在第 i 个样本个案上的值。为了更好地进行下面的讲解，可将 $x^{(i)}$——第 i 个个案所对应的协变量值的矢量——看作(列)矢量 k。

误差方差的估计量通常由 $\hat{\sigma}^2 = \mathrm{RSS}/(n-k)$ 表示，在这里，RSS 表示剩余方差，而 k 则是拟合模型中预测变量的个数(包括常数项)。将预测变量值的 $n\times k$ 矩阵表示为 X[因此第 i 列为 $x^{(i)}$，即第 i 个个案所对应的协变量值]，回归系数矩阵中的最小二乘估计量 $\hat{\beta}$ 的联合分布是多元正态的，其均值为真实值 β，

协方差矩阵为 $\hat{\sigma}^2(X'X)^{-1}$。因此，一个单独的系数估计量 $\hat{\beta}_j$ 服从正态分布，其均值为真实值 β_j，方差为 $\delta_j\sigma^2$，这里 δ_j 表示矩阵 $(X'X)^{-1}$ 的第 j 个对角线元素(diagonal entry)。这样，我们通过 $\delta_j\hat{\sigma}^2$ 估计方差 $\hat{\beta}_j$。

当然，我们通过对估计量取平方根而估计其标准差，并将之看作 $\hat{\beta}_j$ 的标准误(表示为 $s_{\hat{\beta}_j}$)。关于误差分布假设的结果是，量 $(\hat{\beta}_j-\beta_j)/s_{\hat{\beta}_j}$ 服从自由度为 $n-k$ 的学生 t 分布。这允许我们建立 β_j 的标准的 $100(1-\alpha)\%$ 置信区间，表示为 $\hat{\beta}_j \pm t_{\alpha/2}s_{\hat{\beta}_j}$，同样允许在 α 水平下通过拒绝虚无假设 $H_0:\beta_j=0$，如果 $|\hat{\beta}_j/s_{\hat{\beta}_j}|>t_{\alpha/2}$，进而检验第 j 个协变量是否对因变量有显著影响。

而且，这些结果对于大样本同样有效，甚至当我们放宽标准误的正态性假设时。如果是那样的话，量 $(\hat{\beta}_j-\beta_j)/s_{\hat{\beta}_j}$ 接近于标准正态分布。因此，上面描述的假设检验和置信区间，还可以用 $z_{\alpha/2}$ 替代 t 分布中关键的上 $\alpha/2$ 点，$z_{\alpha/2}$ 表示标准正态分布中关键的上 $\alpha/2$ 点。

表 4.1 展示了线性回归模型的拟合值结果，在这里，收入是两个预测变量(ED 和 $WHITE$)的函数。下面给出了估计系数，并在括号里给出了它们的标准误。例如，对于 ED，它的标准误估计为 98。$WHITE$ 的系数同样拥有一个小的标准误：777。

表 4.1 收入线性回归估计值的渐近标准误

变　量	收　入
ED	6294**
	(98)
WHITE	11317**
	(777)
决定系数	0.16

注：** $p<0.01$。

QRM 的标准误和置信区间

我们希望对表达式为 $Q^{(p)}(y_i \mid x^{(i)}=\sum_{j=1}^{k}\beta_j^{(p)}x_j^{(i)})$ 的 QRM 的系数 $\beta^{(p)}$ 进行

推论。正如第 3 章提到的，与这个模型等同的表达式可写为 $y_i = \sum_{j=1}^{k}\beta_j^{(p)}x_j^{(i)} + \varepsilon_i^{(p)}$，这里，$\varepsilon_i^{(p)}$ 拥有第 p 分位数为 0 的一般分布。正如 LRM 的情况一样，系数$\beta_j^{(p)}$的推论是在计算 $\hat{\beta}_j^{(p)}$ 的标准误 $s_{\hat{\beta}_j^{(p)}}$ 的基础上，以置信区间或者假设检验的形式进行。这个标准误具有一个特性，即量 $(\hat{\beta}_j^{(p)} - \beta_j^{(p)})/s_{\hat{\beta}_j^{(p)}}$ 接近标准正态分布。

正如第三章提到的，QRM 的标准误在 i. i. d 模型中是比较简单和容易描述的。因此，$\hat{\beta}_j^{(p)}$ 的渐近协方差矩阵可以写成：

$$\sum_{\hat{\beta}^{(p)}} = \frac{p(1-p)}{n} \cdot \frac{1}{f_{\varepsilon^{(p)}}(0)^2}(X'X)^{-1} \qquad [4.1]$$

在方程 4.1 中出现的项式 $f_{\varepsilon^{(p)}}(0)$代表着从误差分布的第 p 分位数估计得到的误差项$\varepsilon^{(p)}$的概率密度。[8]正如在 LRM 中，协方差矩阵是矩阵$(X'X)^{-1}$的纯量倍数(scalar multiple)。然而，在 QRM 中，乘数 $\frac{p(1-p)}{n} \cdot \frac{1}{f_{\varepsilon^{(p)}}(0)^2}$ 代表基于(单变量)样本 $\varepsilon_1^{(p)}, \cdots, \varepsilon_n^{(p)}$ 上的样本分位数的渐近方差。在方程 4.1 中出现的密度项是未知的，如同在单变量样本中一样需要被估计，而且在第二章描述的估计对应项的步骤可轻易地应用到现在的情境中。量 $\frac{1}{f_{\varepsilon^{(p)}}} = \frac{d}{dp}Q^{(p)}(\varepsilon^{(p)})$ 可通过不同的商 $\frac{1}{2h}[\hat{Q}^{(p)}(p+h) - \hat{Q}^{(p)}(p-h)]$ 估计得到，在这里，样本分位数 $\hat{Q}(p\pm h)$ 是以拟合 QRM 模型中的残差 $\hat{\varepsilon}_i^{(p)} = y_i \sum_{j=1}^{k}\beta_j^{(p)}x_j^{(i)}$，$i = 1, \cdots, n$为基础的。对 h 的选择比较微妙，凯恩克(2005)提出了一些可以选择的方法。

处理非 i. i. d 样本的情况将更加复杂。在这种情况下，$\varepsilon_i^{(p)}$不再具有共同的分布，但所有这些分布依然拥有为 0 的第 p 分位数。为了处理这些非共同的分布，有必要引入矩阵 $X'X$ 的加权版本(下面的 D_1)。

所有这些在 QRM 中获取渐近标准误的分析方法，始于凯恩克在书中提到的一般结论(Koenker, 2005)，该书提出了一种适合系数估计值 $\hat{\beta}_j^{(p)}$联合分布的多变量标准渐近法(multivariate normal approximation)。这种分布的均值由真

实系数和协方差矩阵组成，这一矩阵的表达式为 $\sum_{\hat{\beta}^{(p)}} = \frac{p(1-p)}{n} D_1^{-1} D_0 D_1^{-1}$，又有：

$$D_0 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x^{(i)t} x^i，和\ D_1 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} w_i x^{(i)t} x^i \qquad [4.2]$$

这里 $x^{(i)}$ 表示 $1\times k$ 维度下的 X 的第 i 列。而 D_0 和 D_1 指的是 $k\times k$ 矩阵。加权值为 $w_i = f_{\varepsilon_i^{(p)}}(0)$，其中概率密度函数 $\varepsilon_i^{(p)}$ 的估值为 0[这就是 $\varepsilon_i^{(p)}$ 的第 p 条件分位数]。这样，我们可以求出表达式为 $\widetilde{X}'\widetilde{X}$ 的 D_1 的总和，这里 $\widetilde{X}$ 是从 X 的第 i 列乘以 $\sqrt{w_i}$ 而得到。在方程 4.1 趋向于正定矩阵 D_i 的情况下可以给出适度宽松的条件。正如在 i. i. d 情况中一样，我们看到：条件密度函数之上的 $\hat{\beta}^{(p)}$ 渐近分布是基于从感兴趣的分位数估值得到的。然而，由于 $\varepsilon_i^{(p)}$ 不服从同分布，这些项式随着 i 的不同而不同，于是产生了不同的权重。由于密度函数是未知的，它成为估计出现在方程 4.2 中的权重 w_i 的必须条件。计算这些权重的估计值 $\hat{w}_i$ 的两种方法在凯恩克的书中有所提及。不管应用的方法是哪种，$\hat{\beta}^{(p)}$ 的协方差矩阵都被估算为 $\hat{\Sigma} = \frac{p(1-p)}{n} \hat{D}_1^{-1} \hat{D}_0 \hat{D}_1^{-1}$，在这里：

$$\hat{D}_0 = \frac{1}{n} \sum_{i=1}^{n} x^{(i)t} x^i，和\ \hat{D}_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i x^{(i)t} x^i \qquad [4.3]$$

单独的系数估计量 $\hat{\beta}_t^{(p)}$ 的估计标准误可通过估计协方差矩阵 $\hat{\Sigma}$ 的相应对角线元素(corresponding diagonal element)的平方根获得。如在 i. i. d 情况中，现在我们可以检验关于协变量对因变量影响效应的假设，并且获得分位数回归系数的置信区间。

表 4.2 展示了二元变量 QRM 中第 0.05 和第 0.95 的收入分位数估计值的渐近和自举标准误。虽然渐近和自举的标准误存在一定的差异，但它们关于 *ED* 和 *WHITE* 效应的结论是相同的。在第 0.05 分位数上，*ED* 的点估计为 1130 美元，标准误为 36 美元。而在第 0.95 分位数上，相对应的数字分别为 9575 美元和 605 美元。但在第 0.05 分位数上，*WHITE* 的系数等于 3197 美元，标准误为 359 美元；在第 0.95 分位数上则为 17484 美元和 2895 美元。我们

还可以用标准误计算置信区间。

表 4.2 收入分位数回归模型—渐近和 500 次重抽样自举标准误
(500 Resample Bootstrap Standard Errors)

变　量	p	
	0.05	0.95
ED	1130	9575
	(36)	(605)
	[80]	[268]
WHITE	3197	17484
	(359)	(2895)
	[265]	[2280]

注:小括号内为渐近标准误;中括号内为自举标准误。

表 4.2 表明:两个极端的分位数上的 *ED* 和 *WHITE* 的正向效应在统计上是显著的。然而,协变量效应在不同的分位数上是否存在显著差异,还需要进一步检验。这些检验需要不同分位数上的系数的协方差矩阵。正如我们上面讨论过的,估计 QRM 的误差方差要比 LRM 复杂得多;因此,从多元 QRM 得到的系数协方差将会十分复杂,以至于在实践中不可能得到封闭解。这样,我们需要替代性的方法,来估计不同分位数上的系数协方差,这个话题我们将在下一个部分讨论。

关于渐近标准误,更重要的担心是 i. i. d 假设通常难以成立。经常观察到的偏态和离群值使得误差分布背离 i. i. d 假设。我们发现标准的大样本近似法对于 i. i. d 误差假设的微小偏离都是十分敏感的。这样,基于强参数假设之上的渐近程序不适合用于完成假设检验和估计置信区间(Koenker, 1994)。不需要 i. i. d 假设的替代性方法是更加稳健和实用的(例如:Kocherginsky, He & Mu, 2005)。为了获得稳健的结果,就要求不论因变量的概率密度函数的形式是什么,其统计方法都是适用的,而且它的误差也是令人满意的。换言之,这种替代方法对因变量的分布没有前提假设。一个好的选择便是自举法。

QRM 的自举法(Bootstrap Method)

可代替在之前章节中讨论的渐近法的便是自举法。自举法是一种蒙特卡洛(Monte-Carlo)方法，它用于估计参数估计值的抽样分布，这些参数估计值的计算依赖于来自某种分布的大小为 n 的样本。当一般的蒙特卡洛模拟法被用于模拟抽样分布时，需要假设总体分布是已知的，大小为 n 的样本是从该总体分布中抽取的，而且每一个样本都被用于计算参数估计值。这些从计算得来的参数估计值的经验分布便被用作模拟我们想要的抽样分布。特别的，估计值的标准误可以通过参数估计值的样本的标准差计算得到。

由埃弗隆(Efron, 1979)引入的自举法不同于一般的蒙特卡洛模拟法。与从假定的分布中抽取样本不同的是，我们从实际观察得到的数据中采用放回抽样方法抽取大小为 n 的样本。重新抽样的数量(用 M 表示)在估计标准差时通常位于 50 到 200 之间，而在估计置信区间时，则位于 500 到 2000 之间。尽管每一个重新抽样有着与初始样本相同的要素数量，但它可以多次抽取某些初始数据点，同时排除其他一些数据点。因此，每一个再抽样样本都随机地偏离初始样本。

为了举例说明自举法，我们来考虑在样本 $y_1, \cdots, y_n$ 的第 25 百分位数 $\hat{Q}^{(0.25)}$ 样本基础上估计总体第 25 百分位数 $Q^{(0.25)}$ 的情况。我们想要计算估计值的标准误。一种方法是使用第 2 章提到的大样本模拟 $\hat{Q}^{(0.25)}$ 的方差。这将 $\sqrt{\frac{p(1-p)}{nf(Q^{(p)})^2}}=\sqrt{\frac{(1/4)(3/4)}{nf(Q^{(p)})^2}}=\frac{\sqrt{3}}{4\sqrt{nf}(Q^{(p)})}$ 作为 $\hat{Q}^{(0.25)}$ 的标准差的近似值，在这里 f 表示总体密度函数。由于这个密度是未知的，所以我们必须估计它，并且正如在本章开始部分提到的，我们可以通过公式 $(\hat{Q}^{(0.25+h)}-\hat{Q}^{(0.25-h)})/(2h)$ 和选择适当的常数 h 来估计项式 $1/f(\hat{Q}^{(0.25)})$。

同样解决这一问题的自举法在某种程度上更加直接：我们采用放回抽样方法，从初始样本中抽取大小为 n 的大量样本。这些样本被称为自举样本(bootstrap sample)。对于第 m 个自举样本 $\tilde{y}_1^{(m)}, \cdots, \tilde{y}_n^{(m)}$，我们要计算值 $\hat{Q}_m^{(0.25)}$。重

复大数 M(50 至 200)次抽样，将产生样本 $\hat{Q}_m^{(0.25)}$，$m=1, \cdots, M$，我们将其视为从抽样分布$\hat{Q}^{(0.25)}$中抽取而来的。然后，我们使用 $\hat{Q}_m^{(0.25)}$，$m=1, \cdots, M$ 的标准差 s_{boot}，来估计我们想要的标准差。

自举估计同样用来获得目标总体第 25 百分位数的近似置信区间。有各种各样的方法可以达到这一目的。其中一种是利用样本的初始估计值$\hat{Q}^{(0.25)}$及它的标准误 s_{boot}，$100(1-\alpha)\%$ 来计算置信区间的常态近似值 $\hat{Q}^{(0.25)} \pm z_{\alpha/2} s_{boot}$。

另一种替代方法是利用自举估计样本的经验分位数。对于自举的 95%置信区间，我们将区间的端点看作样本自举估计值的第 0.25 和第 0.975 分位数。更具体的，如果我们将自举估计值$\hat{Q}_1^{(0.25)}$，$\cdots$，$\hat{Q}_{1\,000}^{(0.25)}$从小到大地排列数字顺序 $\hat{Q}_1^{(0.25)}$，$\cdots$，$\hat{Q}_{1\,000}^{(0.25)}$，我们的置信区间便是$[\hat{Q}_{50}^{(0.25)}, \hat{Q}_{951}^{(0.25)}]$。其他任何覆盖概率的置信区间都可以使用相似的方法建立。

将这一思路扩展至 QRM，我们希望估计分位数回归的参数估计值 $\beta^{(p)} = (\beta_1^{(p)}, \cdots, \beta_k^{(p)})$ 的标准误，这些参数估计值是基于包括样本协变量—因变量数据对 (x_i, y_i)，$i=1, \cdots, n$ 的数据而计算得来的。这种(x, y)数据对的自举法，指的是通过对这些数据对的放回抽样而得到的大小为 n 的自举样本，数据对即微观单位(包括 x，y 数据的个案)。样本中一个数据对的重复次数是通过他们的重复率而计算的，因此，一个出现了 k 次的数据对被抽取的概率将增加 k 倍。

每一个自举样本产生一个参数估计值，而我们通过 M 自举估计值可以计算特定系数估计值$\hat{\beta}_i^{(p)}$的标准误 s_{boot}。这种自举估计可以通过不同方式用于计算单个分位数回归参数 $\beta_i^{(p)}$ 的置信区间。其中一种便是利用标准误估计和正态渐近法：$\hat{\beta}_i^{(p)} \pm z_{\alpha/2} s_{boot}$。或者，我们可以在样本分位数的基础上计算置信区间。例如 $\hat{\beta}_i^{(p)}$ 的 95%置信区间从样本的第 2.5 百分位数延续至第 97.5 百分位数，这一样本包括 M 的自举估计值$\hat{\beta}_m^p$。

例如，对基于 19 个等距分位数 ($p=0.05, \cdots, 0.95$) 的多元 QRM 可以从总体上考虑。我们可以估计 19 个模型中所有可能的分位数回归系数之间的协方差。例如：当正在拟合的模型包括截距参数$\hat{\beta}_1^p$和与两个协变量$\hat{\beta}_2^p$和$\hat{\beta}_3^p$相对应

的系数时，我们有 3×19 = 57 个被估计的系数，产生 57×57 协方差矩阵。这一矩阵不仅提供了在每一个分位数[如，Var($\hat{\beta}_1^{0.05}$)和 Var($\hat{\beta}_1^{(0.50)}$)]上各个协变量的系数方差，而且还提供同一协变量[如，Cov($\hat{\beta}_1^{0.05}$)，Var($\hat{\beta}_1^{0.50}$)]在不同分位数上的估计值的协方差。

通过对方差和协方差的估计，我们可以通过瓦尔德(Wald)值进行假设检验，以验证同一协变量但不同分位数 p 和 q 相对应的一对系数 $\beta_i^{(p)}$ 和 $\beta_i^{(q)}$ 是否相等。

$$Wald\text{ 值} = \frac{(\hat{\beta}_j^{(p)} - \hat{\beta}_j^{(q)})^2}{\hat{\sigma}^2_{\hat{\beta}_j^{(p)} - \hat{\beta}_j^{(q)}}} \quad [4.4]$$

分母中的 $\hat{\sigma}^2_{\hat{\beta}_j^{(p)} - \hat{\beta}_j^{(q)}}$ 是 $\hat{\beta}_j^{(p)} - \hat{\beta}_j^{(q)}$ 的差的估计方差，可以通过下面的等式和替代等式右端的估计方差和协方差而获得：

$$\mathrm{Var}(\hat{\beta}_j^{(p)} - \hat{\beta}_j^{(q)}) = \mathrm{Var}(\hat{\beta}_j^{(p)}) + \mathrm{Var}(\hat{\beta}_j^{(q)}) - 2\mathrm{Cov}(\hat{\beta}_j^{(p)}, \hat{\beta}_j^{(q)}) \quad [4.5]$$

在虚无假设下，瓦尔德值服从自由度为 1 的卡方分布(χ^2)。

更一般的，我们可以检验不同分位数上的多元系数是否相等。例如，假设我们拥有两个协变量和模型中的截距部分，我们希望检验第 p 和第 q 条件分位数函数是否可以彼此转化，即：

$$H_0: \beta_2^{(p)} = \beta_2^{(q)} \text{ 和 } \beta_3^{(p)} = \beta_3^{(q)} \text{ 与 } H_a: \beta_2^{(p)} \neq \beta_2^{(q)} \text{ 或者 } \beta_3^{(p)} \neq \beta_3^{(q)}$$

去除了截距部分。进行此类检验的瓦尔德(Wald)值可以被描述如下：首先，我们利用估计的协方差来获得$\hat{\beta}^{(p)} - \hat{\beta}^{(q)}$ 的一个估计协方差矩阵 $\hat{\Sigma}_{\hat{\beta}^{(p)} - \hat{\beta}^{(q)}}$，其表达式为 $\hat{\Sigma}_{\hat{\beta}^{(p)} - \hat{\beta}^{(q)}} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} \end{bmatrix}$，这个式子可以通过将估计的方差和协方差替换为下面的式子而求解：

$$\hat{\sigma}_{11} = \mathrm{Var}(\hat{\beta}_1^{(p)} - \hat{\beta}_1^{(q)}) = \mathrm{Var}(\hat{\beta}_1^{(p)}) + \mathrm{Var}(\hat{\beta}_1^{(q)}) - 2\mathrm{Cov}(\hat{\beta}_1^{(p)}, \hat{\beta}_1^{(q)})$$

$$\hat{\sigma}_{12} = \hat{\sigma}_{21} = \mathrm{Cov}(\hat{\beta}_1^{(p)}, \hat{\beta}_2^{(p)}) + \mathrm{Cov}(\hat{\beta}_1^{(q)}, \hat{\beta}_2^{(q)}) - \mathrm{Cov}(\hat{\beta}_1^{(p)}, \hat{\beta}_2^{(q)}) - \mathrm{Cov}(\hat{\beta}_1^{(q)}, \hat{\beta}_2^{(p)})$$

$$\hat{\sigma}_{22} = \mathrm{Var}(\hat{\beta}_2^{(p)} - \hat{\beta}_2^{(q)}) = \mathrm{Var}(\hat{\beta}_2^{(p)}) + \mathrm{Var}(\hat{\beta}_2^{(q)}) - \mathrm{Cov}(\hat{\beta}_2^{(p)}, \hat{\beta}_2^{(q)})$$

下面我们计算这一检验值：

$$W = \begin{bmatrix} \hat{\beta}_1^{(p)} & \hat{\beta}_1^{(q)} \\ \hat{\beta}_2^{(p)} & \hat{\beta}_2^{(q)} \end{bmatrix}^t \hat{\Sigma}^{-1}_{\hat{\beta}^{(p)} - \hat{\beta}^{(q)}} \begin{bmatrix} \hat{\beta}_1^{(p)} & \hat{\beta}_1^{(q)} \\ \hat{\beta}_2^{(p)} & \hat{\beta}_2^{(q)} \end{bmatrix}$$

这个式子在虚无假设下，接近于自由度为 2 的卡方分布(χ^2)。

Stata 通过 bsqreg 命令，可实现单个 QRM 的自举程序，而对于多元 QRM，则采用 sqreg 命令。从 sqreg 命令计算出的估计值和那些通过 bsqreg 命令得到的个别结果是一样的，但是 sqreg 命令会提供整个协方差矩阵。sqreg 的实用性有助于研究者检验不同分位数上的系数是否相等。依靠计算机技术的发展，多数研究者得以使用自举法。例如，在超过 2 万户家庭的收入数据中的 500 个重新抽样样本，中位数上的二元 QRM 的协方差估计任务，Stata(版本 9.2)通过 64 位、1.6G 赫兹的处理器大概要用 8 分钟来完成。而拥有 500 个复制样本的 19 个分位数进行类似估计则需要花 2 个小时。

QRM 的拟合优度

在线性回归模型中，拟合优度由 R^2(即决定系数)测量：

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad [4.6]$$

在第二项表达式中的分子代表观察值 y_i 与相应的模型拟合值 $\hat{y}_i$ 的距离平方和。另一方面，分母则是观察值和拟合值(只从模型中的截距部分计算得来)的距离平方和。因此，R^2 可被理解为在模型中因变量的差异可由预测变量解释的比例。这个数值落在 0 和 1 之间，R^2 值越大，表示模型拟合越好。

对于分位数回归模型，我们可以轻松地发展出类似于 R^2 的统计值。由于线性回归模型拟合是基于最小二乘法的，而分位数回归模型则以最小化加权距

离总和（$\sum_{i=1}^{n} d_p(y_i, \hat{y}_i)$）为基础，正如第 3 章至第 5 章提到的——根据 $y_i > \hat{y}_i$ 还是 $y_i < \hat{y}_i$ 而确定不同的权重——我们依据符合这一标准的方式来测量拟合优度。凯恩克和马沙杜(Koenker & Machado，1999)建议通过比较模型中的加权距离总和与截距参数的总和来测量拟合优度。让 $V^1(p)$ 表示第 p 分位数回归完整模型的加权距离总和，让 $V^0(p)$ 表示只包括常数项的模型的加权距离总和。以单一协变量模型为例，我们有：

$$V^1(p) = \sum_{i=1}^{n} d_p(y_i, \hat{y}_i) = \sum_{y_i \geqslant \beta_0^{(p)} + \beta_1^{(p)} x_i} p \mid y_i - \beta_0^{(p)} - \beta_1^{(p)} x_i \mid$$
$$+ \sum_{y_i < \beta_0^{(p)} + \beta_1^{(p)} x_i} (1-p) \mid y_i - \beta_0^{(p)} - \beta_1^{(p)} x_i \mid$$

和

$$V^0(p) = \sum_{i=1}^{n} d_p(y_i, \hat{Q}^{(p)}) = \sum_{y_i \geqslant \hat{y}} p \mid y_i - \hat{Q}^{(p)} \mid + \sum_{y_i < \hat{y}} (1-p) \mid y_i - \hat{Q}^{(p)} \mid$$

对于只包含常数项的模型，拟合的常数正是样本 $y_1, \cdots, y_n$ 的第 p 分位数 $\hat{Q}^{(p)}$。这样，拟合优度便被定义为：

$$R_p = 1 - \frac{V^1(p)}{V^0(p)} \tag{4.7}$$

由于 $V^0(p)$ 和 $V^1(p)$ 是非负的，R_p 的最大值为 1。同样，因为完整拟合模型中令加权距离总和最小化，$V^1(p)$ 永远不会大于 $V^0(p)$，所以 R_p 大于或等于 0。因此，R_p 的取值范围在 0 和 1 之间，大的 R_p 值表示更好的模型拟合。方程 4.7 是 QRM 在 p 上的拟合优度的局部测量。对于所有分布的 QRM 的全面评估，要求对 R_p 进行整体检验。

上面定义的 R_p 允许我们对超出截距项的任意协变量拟合模型与只包括截距项的模型进行比较。这是由凯恩克和马沙杜(1999)为嵌套模型引进的比较拟合优度的受限形式。通过明显的扩展，对于特定模型拟合度的改善程度，可以通过模型的较大受限形式来测量。由此得到的结果被称为相对 $R_{(p)}$ 值。让 $V^2(p)$ 表示限制较小的第 p 分位数回归模型的加权距离总和，而让 $V^1(p)$ 表示限制较大的第 p 分位数回归模型的加权距离总和。而相对 $R_{(p)}$ 值可表示为：

$$\text{相对 } R_{(p)} = 1 - \frac{V^2(p)}{V^1(p)} \qquad [4.8]$$

下面我们用收入的例子进行讲解。我们在 19 个等距分位数上对收入拟合一个二元变量 QRM(教育和种族)和一个一元变量 QRM(只有教育)。表 4.3 中的数值分别代表常数模型和完整模型的拟合优度(见图 4.1)。Stata 通过方程 4.7 提供了拟合优度的测量值,并称之为伪(pseudo)R^2 值,以便与 LRM 的常规 R^2 相区分。

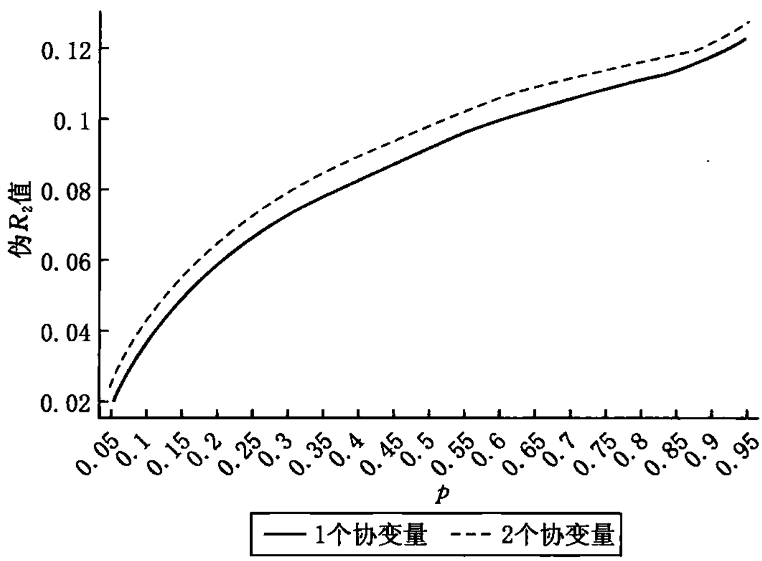


图 4.1　分位回归模型的拟合优度:一个协变量模型嵌套于两个协变量模型

表 4.3 列出了二元协变量模型的拟合优度。在低尾处的收入拟合优度要低于上尾处。收入分布中 19 个分位数上的平均 $R_{(p)}$ 为 0.0913。而一元协变量模型是嵌套在二元协变量模型中的,它在 19 个分位数上的平均 $R_{(p)}$ 为 0.0857。这些模型的 $R_{(p)}$ 表示将种族作为解释变量会改善模型的拟合。对比于一元协变量模型的 $R_{(p)}$,二元协变量模型的 $R_{(p)}$ 只有适度的增加,所以,教育拥有主要的解释力。对于添加种族是否显著改善模型的正式检验方法是 t 比率(t-ratio)。对于一组解释变量的正式检验方法超出了本章的范围,感兴趣的读者可参考凯恩克和马沙杜的书(Koenker and Machado, 1999)。

表 4.3 收入分位数回归模型的拟合优度

模型	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
二元变量—收入	0.0254	0.0441	0.0557	0.0652	0.0726	0.0793	0.0847	0.0897	0.0943	0.0985
一元变量—收入	0.0204	0.0381	0.0496	0.0591	0.0666	0.0732	0.0784	0.0834	0.0881	0.0922

模型	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	均值
二元变量—收入	0.1025	0.1059	0.1092	0.1120	0.1141	0.1162	0.1179	0.1208	0.1271	0.0913
一元变量—收入	0.0963	0.0998	0.1033	0.1064	0.1092	0.1112	0.1131	0.1169	0.1230	0.0857

注:二元变量模型包括教育和种族;一元变量模型只包括教育。表格内数字为 *R*—QRM 的拟合优度测量值。

小结

本章讨论了分位数回归模型的推论统计。QRM 系数的渐近推论(标准误和置信区间)类似于 LRM 系数的推论方法,只要经过必要的修改,便可以正确地估计误差方差。假设在社会科学研究中因变量常常出现偏态分布,那么关于渐近推论的假定是有问题的,所以需要一种替代性的推论方法。自举法出色地解决了这一问题。本章介绍了 QRM 系数的自举程序。自举抽样的理念是相当直观的,并且伴随着计算机技术的发展而变得十分实用。

另外,本章简要讨论了与 LRM 相似的 QRM 拟合优度。QRM 拟合优度的测量值 $R_{(p)}$ 解释了在特定分位数方程中每个观察值所拥有的适当权重。$R_{(p)}$ 是容易理解的,并且对它的解释类似于大家熟悉的 LRM 的 R^2。

第 5 章 | 分位数回归估计值的解释

在本章中，我们讨论对分位数回归估计值的解释。首先，我们需要解释特定分位数的分位数回归拟合。中位数回归中的分位数可用来追踪位置的变化。其他具体的回归分位数，例如，第 0.05 和第 0.95 分位数，可以用作评估一个协变量如何预测因变量的条件非中心位置和形状的变化。我们同样关注序列回归分位数的一般情况，它可以反映因变量分布在形状上更微妙的变化。

我们以 LRM 估计值的解释方法开始，接着在收入不平等的例子中解释 QRM 的估计值。在这方面，我们将论证 QRM 方法优于 LRM 方法的两个主要优势：它帮助我们模型化因变量分布中的非中心条件分位数及其形状变化。在第 3 章中，我们通过同一个收入样本解释各种各样的方法，但现在我们同时考虑教育和种族的影响。纵观全章，我们集中关注初始单位因变量的分析。对单调转换因变量估计值的解释和对因变量初始单位含义的理解将在第 6 章讨论。

参照与比较

为了帮助读者理解分位数回归的估计值，我们采用参照与比较（reference and comparison）的概念和一些关于效应量化的更一般的概念。参照是传统的回归术语，比较则表示在回归中协变量每单位增量的效应。[9]

在多数情况下，我们的兴趣在于进行组别间的比较。例如，我们希望比较受教育年限为 11 年的个体和那些受教育年限为 12 年的个体。或者，我们对黑人和白人的比较感兴趣。无论如何，我们从一组可能的协变量开始，例如，受教育年限为 11 年的所有黑人，我们将拥有这些特征的子总体设为参照组。然后，

我们以特定方式修改其中一个协变量，例如，将 11 年教育水平改为 12 年，或者将黑人种族换成白人。那么，我们将与修改后的一组协变量相对应的子总体作为比较组。这两个组别比较的主要特征是修改了单个协变量，而保持其他协变量不变。

当我们从一个参照组转换到一个比较组时，检验因变量的分布是如何变化的，有助于量化单个协变量的变化对因变量分布的影响效应。对于 LRM 而言，拟合系数可以理解为估计效应，即因变量分布的均值变化估计值，而均值的变化来源于一个连续型协变量的单位增量，或者虚拟变量值从 0 到 1 的变化。每一种变化都可以理解为参照组和比较组在均值上的估计差异。对 QRM 的解释与之类似，即参照组和比较组在特定分位数上的估计差异；当其他协变量保持不变时，这一估计差异来自一个连续型协变量的单位增量，或者虚拟变量值从 0 到 1 的变化。

条件均值与条件中位数

到目前为止，最容易理解的 QRM 是中位数回归模型(即第 0.5QRM)，它用以表达在特定预测变量下因变量的条件中位数，并且中位数回归模型可作为拟合条件均值的 LRM 的一种替代方法。当两种模型试图模型化因变量分布的中心位置时，它们是适合相互比较的。

收入 LRM 的估计值在 *ED* 变量上是 6314 美元，在 *WHITE* 种族上是 11452 美元。受教育年限每增加一年，在任何固定的教育水平上的收入均值将增加 6314 美元。由于线性假设的存在，对于任何固定教育水平下的家庭而言，条件均值的增加量是相同的。例如，受教育年限每增加 1 年，家庭的平均收入便增加相同的数量，不管户主拥有 9 年或 16 年教育年限。另外，教育年限每增加 1 年的效应，对于黑人和白人是同样的：如果模型中没有放入种族和教育的交互项。就参照组和比较组而言，我们可以说：尽管存在许多不同的参照组/比较组的组合，但是只存在两种可能的效应：单一的种族效应和单一的教育效应。[10]

LRM 包含一个严格的假设：从一组到下一组，收入分布是在尺度和形状维

持不变时发生移动的。特别的，教育的正向效应反映了受教育年限每增加 1 年时，分布向右移动的幅度，而这是表现分布移动的唯一方式。相似的，收入 LRM 中 *WHITE* 的系数表示从黑人的收入分布到白人的收入分布存在一个向右的位置移动，同样没有改变分布形状：黑人的平均收入比白人的低 11452 美元。

下面我们从 LRM 转到 QRM，集中关注中位数回归的特殊例子，我们需要记住的是：主要的不同在于我们是模型化条件中位数而不是条件均值。正如在第 3 章讨论的，中位数可能是一个更让人满意的测量分布中心位置的方法，尤其是当我们试图模型化条件分布的整体行为时。例如，这些条件分布可能是右偏的，这令它们的均值更多地反映了分布上尾部分的情况而不是中间部分的情况。作为一个具体的例子，位于上尾收入分位数的家庭可能对关于教育影响中位数收入的研究分析有着深远的影响。因此，这种分析可能反映了教育对条件均值的效应远大于对条件中位数的效应。

表 5.1　收入的分位数回归估计值及其渐近标准误

	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ED	1130	1782	2315	2757	3172	3571	3900	4266	4549	4794
	(36)	(41)	(51)	(51)	(60)	(61)	(66)	(73)	(82)	(92)
WHITE	3197	4689	5642	6557	6724	7541	8168	8744	9087	9792
	(359)	(397)	(475)	(455)	(527)	(528)	(561)	(600)	(662)	(727)

	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	
ED	5182	5571	5841	6224	6598	6954	7505	8279	9575	
	(86)	(102)	(107)	(129)	(154)	(150)	(209)	(316)	(605)	
WHITE	10475	11091	11407	11739	12142	12972	13249	14049	17484	
	(664)	(776)	(793)	(926)	(1065)	(988)	(1.299)	(1790)	(2895)	

注：括号内为渐近标准误。

对中位数回归系数的解释类似于对 LRM 的系数解释。表 5.1 给出了各种分位数回归模型的估计系数，包括中位数（第 0.5 分位数）回归。就连续型变量而言，系数估计值可被理解为预测变量每一个单位的变化对应于因变量的中位数的变化。LRM 的线性和无交互项的解释结果可应用在中位数回归模型中。

特别的，教育水平的每一年增量对中位数因变量的效应对于所有种族和教育水平来说是相同的，而且种族变化的效应在所有教育水平上也是相同的。

条件中位数模型中的 ED 系数是 4794 美元，低于条件均值模型的系数。这意味着增加一年的教育会让收入平均增长 6314 美元，但对总体的多数人来说，收入的增加并没有这么可观。相似的，在条件中位数的 WHITE 系数是 9792 美元，低于条件均值模型中相应的系数。

括号里表示的是在 *i. i. d* 假设下的估计值的渐近标准误。如果 *i. i. d* 假设成立，教育对收入中位数的影响效应的标准误是 92 美元，t 比率为 52.1，并且 p 值小于 0.001，这些证据拒绝了"教育对中位数收入没有影响"的虚无假设。WHITE 的系数的标准误为 727 美元，而且在 0.001 水平上具有统计显著性。

其他个别条件分位数的解释

有时候，相对于中心位置，研究者对分布的低尾或上尾部分更感兴趣。关注平等性的教育政策重视提高落后学生的考试成绩。在 2000 年，39%的八年级学生的学科成绩低于基础水平。因此，对于教育研究者而言，第 0.39 分位数比均值或中位数更有意义。福利政策以低收入群体为对象。如果全国的贫穷率为 11%，那么第 0.11 收入分位数和低于这一水平的分位数对于福利研究者来说，比中位数或均值更有意义。研究者发现工会成员身份对位于收入分布低端的个案比对达到均值水平的个案有更大的回报(Chamberlain, 1994)。另一方面，对于收入排在总体前 10%的人群，在高名望的私人大学接受教育的情况更加普遍。研究有名望的高等教育的得益时，通常关注第 90 或更高的收入分位数。

在表 5.1 中，19 个分位数的 QRM 拟合系数可以用来检验教育和种族在不同收入分位数上的效应。[11]为了告知福利政策制定者，我们检验条件收入模型中第 0.1 和第 0.05 分位数上的教育和种族系数。我们发现：在第 0.1 分位数上，受教育年限每增加 1 年，将使收入增长 1782 美元，在 0.05 分位数上，这一增量为 1130 美元；在第 0.1 分位数上，黑人和白人的差距为 4689 美元，在 0.05 分

位数上，则为 3197 美元。对右端分布的教育回报率感兴趣的研究者，可以观察在第 0.9 和第 0.95 分位数上的教育估计值。第 0.95 分位数的系数为 9575 美元，远高于在第 0.90 分位数上的 8279 美元，这意味着有名望的高等教育对收入的贡献是不一样的。在 *i.i.d* 假设下，渐近的标准误表明教育效应和种族效应在非中心的分位数上是显著的。

因为 *i.i.d* 是一个非常严格的假设，它假定因变量没有发生形状变化，我们应该采用更加灵活的标准误估计方法，如自举法。表 5.2 列出了基于 500 次再抽样自举程序上的二元协变量的参数点估计和标准误。自举的点估计与渐近估计相近，但它们在不同分位数上的变化幅度比渐近标准误更小，尤其是 ED 变量(参见图 5.1 和图 5.2)。

表 5.2 收入的分位数回归估计的点估计与标准误(500 次再抽样自举法)

	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ED	1130	1782	2315	2757	3172	3571	3900	4266	4549	4794
	(80)	(89)	(81)	(56)	(149)	(132)	(76)	(98)	(90)	(103)
WHITE	3197	4689	5642	6557	6724	7541	8168	8744	9087	9792
	(265)	(319)	(369)	(380)	(469)	(778)	(477)	(545)	(577)	(624)

	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
ED	5182	5571	5841	6224	6598	6954	7505	8279	9575
	(83)	(103)	(121)	(125)	(154)	(151)	(141)	(216)	(268)
WHITE	10475	11091	11407	11739	12142	12972	13249	14049	17484
	(589)	(715)	(803)	(769)	(1.041)	(929)	(1350)	(1753)	(2280)

注：括号内为渐近标准误。

不同分位数系数的等值检验

当多元分位数回归被估计时，我们需要检验那些明显的差异是否在统计上显著。为完成这样的检验，需要估计横贯分位数估计值的协方差矩阵。对协方差矩阵的估计可借助自举法来实现，以允许灵活的误差，并为十分复杂的渐近线公式提供数学解决方法。

表 5.3 给出了点估计值、自举标准误和 p 值，这些数值有助于检验第 p 分位数估计值与中位数估计值、第 $(1-p)$ 分位数估计值及第 $(p+0.05)(p\leqslant 0.5)$ 分位数估计值是否等价。根据这些情况，自举法给出了比渐近线法更小或更大的标准误。例如，在中位数收入水平上，渐近线法给出的教育点估计是 4794 美元，标准误是 92 美元。而自举法计算的相应数值分别是 4794 美元和 103 美元。但在第 0.05 分位数上，与渐近线法相比，自举法给出了较低准确度的教育估计值：自举法标准误是 80 美元，大于渐近标准误(36 美元)。

表 5.3 不同收入分位数的系数等价检验(500 次再抽样自举法)

分位数/变量	系数	P 值			
		与中位数系数有无差异	与第 $(1-p)$ 分位数系数有无差异	与第 $(p+0.05)$ 分位数系数有无差异	4 个系数有无联合差异
第 0.05 分位数					
ED	1130** (80)	0.0000	0.0000	0.0000	0.0000
WHITE	3197** (265)	0.0000	0.0000	0.0000	0.0000
第 0.10 分位数					
ED	1782** (89)	0.0000	0.0000	0.0000	0.0000
WHITE	4689** (319)	0.0000	0.0000	0.0000	0.0000
第 0.15 分位数					
ED	2315** (81)	0.0000	0.0000	0.0000	0.0000
WHITE	5642** (369)	0.0000	0.0000	0.0018	0.0000
第 0.20 分位数					
ED	2757** (56)	0.0000	0.0000	0.0000	0.0000
WHITE	6657** (380)	0.0000	0.0000	0.4784	0.0000
第 0.25 分位数					
ED	3172** (149)	0.0000	0.0000	0.0000	0.0000

（续表）

分位数/变量	系　数	P 值			
		与中位数系数有无差异	与第（1－p）分位数系数有无差异	与第（p＋0.05）分位数系数有无差异	4个系数有无联合差异
WHITE	6724** (469)	0.0000	0.0000	0.0012	0.0000
第0.30分位数					
ED	3571** (132)	0.0000	0.0000	0.0000	0.0000
WHITE	7541** (778)	0.0000	0.0000	0.0142	0.0000
第0.35分位数					
ED	3900** (76)	0.0000	0.0000	0.0000	0.0000
WHITE	8168** (477)	0.0000	0.0000	0.0035	0.0000
第0.40分位数					
ED	4266** (98)	0.0000	0.0000	0.0000	0.0000
WHITE	8744** (545)	0.0028	0.0008	0.1034	0.0002
第0.45分位数					
ED	4549** (90)	0.0000	0.0000	0.0000	—
WHITE	9087** (577)	0.0243	0.0017	0.0243	—
第0.50分位数					
ED	4794** (103)	—	—	0.0000	—
WHITE	9792** (624)	—	—	0.0361	—

注：括号内为标准误，** $p < 0.01$。

检验不同分位数估计值是否等价的瓦尔德测试的 p 值表明：在我们所选的检验中，教育效应在不同的分位数上是有差异的。这些检验比较了当前估计值

(即第 0.05 分位数)与其他 3 个估计值:中位数估计值、与之对应的另一端估计值(第 0.95 分位数)和邻近的更高分位数估计值(在 0.10 分位数)。与此相反,在我们所选的众多检验中,种族效应不存在统计上的差异。例如,在 0.20 分位数上的白人效应与在 0.25 分位数上的差异并不显著。尤其是在中位数之上的收入分位数的种族效应与相邻的第 ($p+0.05$) 分位数的种族效应在统计上不存在显著差异,这一现象正好与教育相反,当 p 上升时,教育效应逐渐增大。

你同样可以验证另一个虚无假设,即对于相同的协变量而言,多于两个分位数的系数是联合相等的。表 5.3 最后一列展示了相同协变量的四个分位数系数的联合检验结果。瓦尔德检验统计值接近于自由度为 3 的 χ^2 分布。检验结果拒绝了虚无假设,并且论证了四个系数中至少有两个是彼此显著不相等的。

通过 QRM 结果解释形状变化

许多社会科学研究,尤其是关于不平等的研究,不仅需要说明位置变化,而且要考虑形状变化,因为在很大程度上仅仅关注位置会让我们忽略关于组别差异的许多信息。关于形状特征的最重要的两个考虑是尺度(或离散)和偏态。

一个图像的视角

因为我们的兴趣在于预测变量如何改变因变量的分布形状,所以我们通过 QRM 来计算多元分位数的估计值。对形状效应的分析比对位置的分析要复杂得多,所以我们需要一个重要的权衡。一方面,形状分析可利用各个分位数的 QRM 估计值的多重集(multiple sets)来实现,而且它可以揭示比单独使用位置效应分析更多的信息。另一方面,描述这些信息是比较麻烦的,需要额外的功夫。特别的,对一系列分位数(例如:0.05, 0.10, …, 0.90, 0.95)的回归系数的检验是难以处理的,所以,QRM 估计值的图像化视角成为解释 QRM 结果的必要途径。

特定协变量的 QRM 系数反映了协变量的单位变化对因变量分布的分位数

的影响。因此,一系列分位数的系数组可以用来确定协变量的单位增量是如何影响因变量的分布形状的。我们通过图像视角来检验系数,进而凸显影响形状变化的效应。对于一个特定的协变量,我们描绘其系数和置信封闭间(confidence envelope),这里 y 轴表示预测变量的效应$\hat{\beta}^{(p)}$, x 轴表示分位数 p。

图 5.1 提供了关于教育和种族(两者以各自的均值为中心)的收入分位数函数的图像。使用那些估计系数(见表 5.1),我们同样可以描绘拟合常数(constant)的图像。因为协变量以其均值为中心,所以常数(constant)表示在协变量均值上的拟合分位数函数,这可称为特殊设置(typical setting)。特殊设置下的条件分位数函数如果在中位数以下的斜率平缓而在中位数以上的斜率陡峭的话,那么它就是右偏的。

ED 效应可以表示为,当种族效应维持不变,在任一教育水平下,教育年限每增加 1 年时,条件收入分位数的变化。教育效应是显著为正的,因为其置信封闭间并没有穿过 0 线(见厚平行线)。图 5.1a 描述了一条向上倾斜的教育效应曲线:教育年限的单位增加效应对于所有第 p 分位数数值都是正向的,还随着 p 的增加而稳定上升。并且这种增加速度在第 0.8 以上分位数有所提升。

WHITE 效应可以表示为:当教育水平保持不变时,将种族变量从黑人换成白人时条件收入分位数的变化幅度。作为白人的效应是显著为正的,因为 0 线远低于置信封闭间。与黑人相比,图 5.1(b)描述了另一条向上倾斜的白人效应曲线。在第 0.15 分位数之下和第 0.90 分位数之上的斜率比那些位于中间分位数的斜率更加陡峭。

图 5.2 对应于图 5.1,但图 5.2 的置信封闭间是基于自举估计得到的。我们观察到,图 5.2 中的自举置信封闭间相对于图 5.1 的渐近置信封闭间要更加平衡。我们从图 5.1 和图 5.2 可看到一个相似的形状变化模式。

这些图像传达了与教育和种族效应相关的额外信息。首先,教育和种族在影响位置变化的同时,也影响着形状变化。如果只存在位置变化,那么增加 1 年的教育或将种族从黑人换成白人,将令所有的分位数的增量相同,从而导致 $\hat{\beta}^{(p)}-p$ 图像接近一条水平线。然而,我们观察到的则是$\hat{\beta}^{(p)}$随着 p 而单调递增,即当 $p>q$ 时,有 $\hat{\beta}^{(p)}>\hat{\beta}^{(q)}$,并且,这一性质告诉我们:增加 1 年教育或者

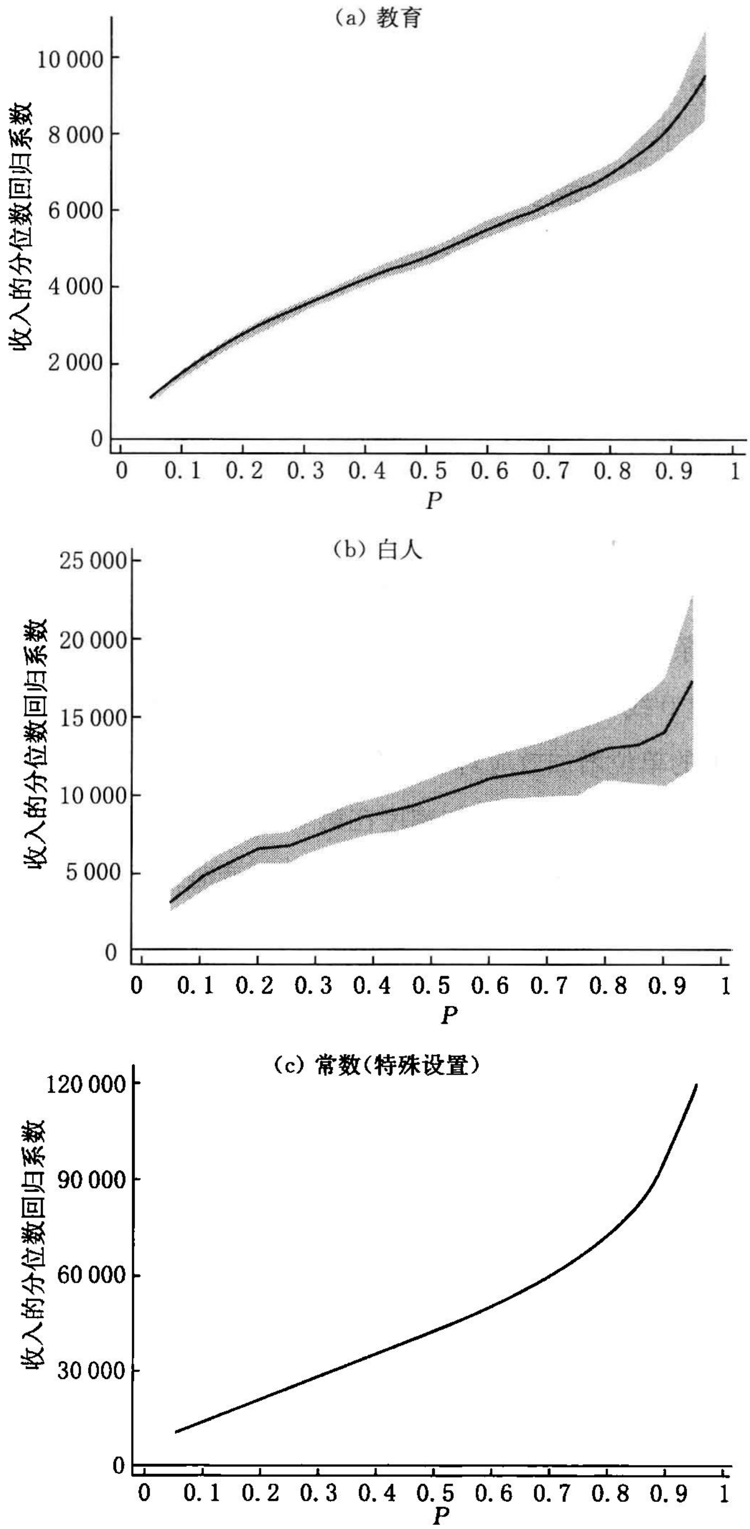


图 5.1 分位回归估计的 95%置信区间渐进线(收入)

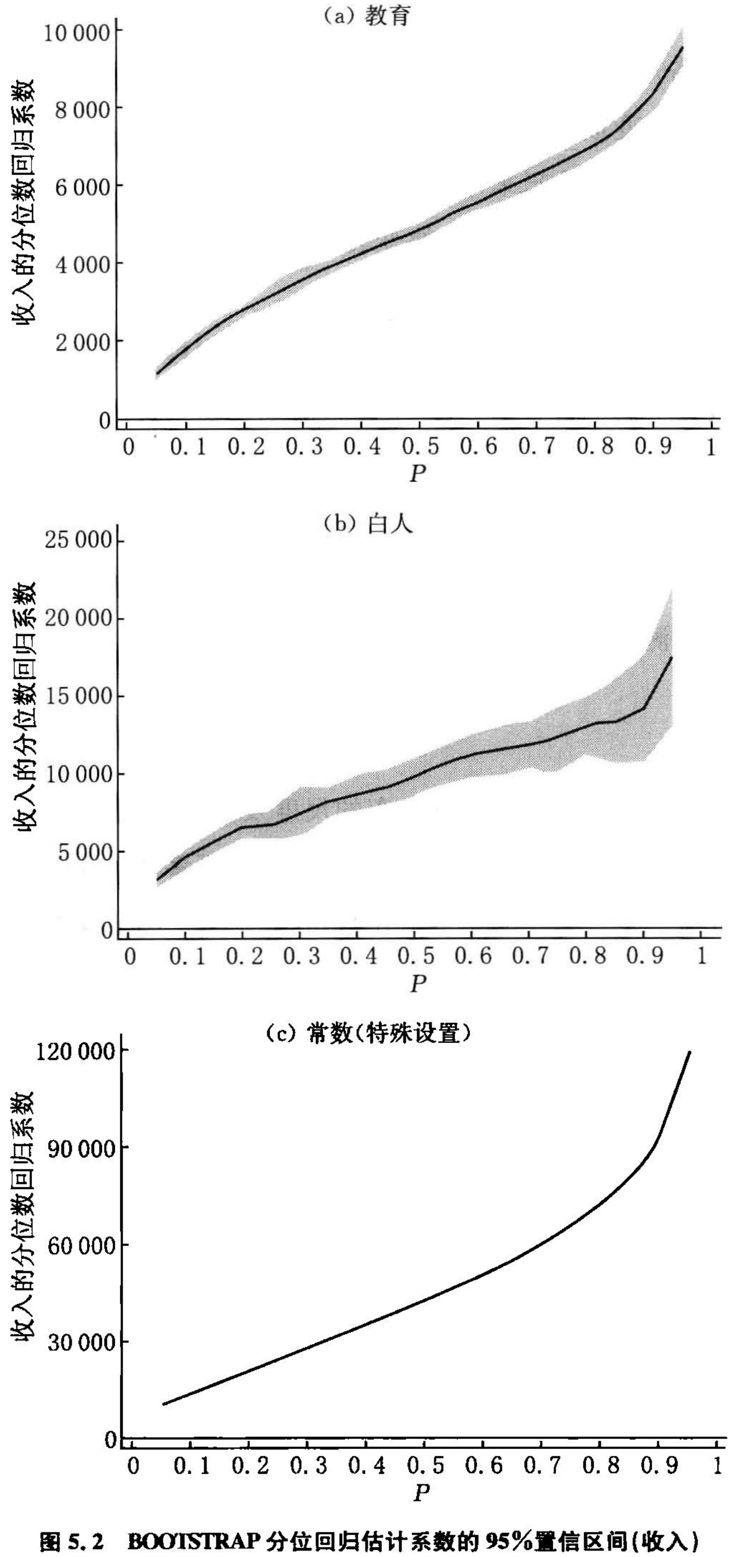


图 5.2 BOOTSTRAP 分位回归估计系数的 95%置信区间(收入)

将种族从黑人变成白人对于高收入阶层的收入效应要大于低收入阶层。这一单调性同样具备尺度效应，因为它表明当 $p<0.5$ 时，有 $\hat{\beta}^{(1-p)}-\hat{\beta}^{(p)}>0$。换言之，将种族从黑人变成白人或者增加一年教育水平会增大因变量的单位尺度。[12]尽管这两个图看起来暗示了真实的变化将比位置和尺度变化更加复杂，但图像并不足以反映偏态的变化，因为偏态需要借助多元分位数来测量。

现在，我们总结一下协变量对因变量效应的图像模式。平行线表明协变量的单位增量仅仅会影响位置变化；向上倾斜曲线表明条件因变量分布的尺度的增大，而向下倾斜曲线则表示尺度的减小。然而，图像的视角不足以展示偏态的变化。图像化提供了关于预测变量的变化是如何产生形状变化的一些证据。我们同样对变化的幅度和变化是否显著感兴趣。我们的下一个目标是从 QRM 估计值中发展出测量两类形状变化的定量方法。

尺度变化

标准差是普遍使用的测量对称分布的尺度及离散程度的方法。然而，对于偏态分布，所选分位数之间的距离提供了比标准差更多的关于离散程度的有用信息。对于在 0 和 0.5 之间的第 p 分位数，我们确定了两个样本分位数：$\hat{Q}^{(1-p)}$（第 $[1-p]$ 分位数）和 $\hat{Q}^{(p)}$（第 p 分位数）。第 p 分位数间距，$IQR^{(p)}=\hat{Q}^{(1-p)}-\hat{Q}^{(p)}$，是测量离散程度的方法。这一数量描述了分布中间 $(1-2p)$ 部分的比例范围。当 $p=0.25$ 时，这一分位数间距便是四分位间距 $IQR^{(0.25)}=\hat{Q}^{(0.75)}-\hat{Q}^{(0.25)}$，提供了分布中间 50% 的比例范围。而 p 的其他值，例如 0.10、0.05、0.025，同样可用以捕捉一个分布两个尾端之间的离散范围。例如，使用 $p=0.10$，第 p 分位数间距表示分布中间 80% 的比例范围。

图 5.3 对比了一个参照组和一个比较组的情况，它们有着相同的中位数 M。确定了 p 的数值后，我们便可以计算参照组的分位数间距 $IQR_R=U_R-L_R$，和比较组的 $IQR_C=U_C-L_C$。然后，我们通过间距的差(difference-in-differences) IQR_C-IQR_R 测量尺度的变化。在此图中，比较组的尺度比参照组的更大，即存在一个正向的尺度变化。

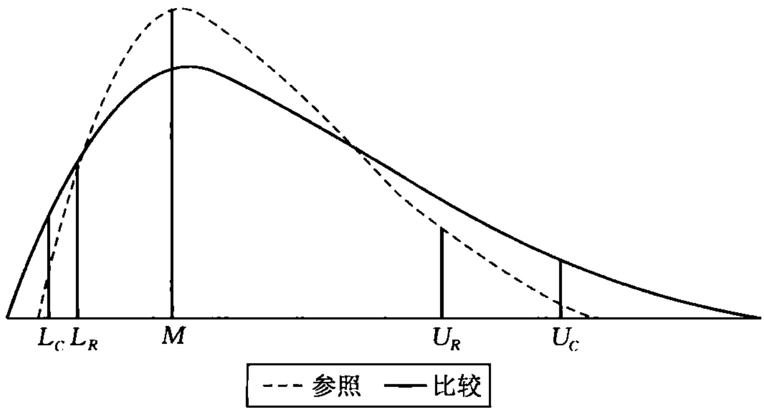


图 5.3 尺度变化

转向我们的应用例子:表 5.4 展示了两个模型下不同教育组家庭收入分布的尺度变化,一种方法使用样本分位数,即分位数直接从两组样本中计算,而第二种方法利用收入 QRM 的协变量拟合系数来实现。对于样本分位数,11 年教育年限组的四分位间距是 26426 美元,12 年教育年限组的则为 34426 美元。12 年教育年限组的样本离散程度为 8000 美元,高于 11 年教育年限组。这一尺度变化可以通过计算两组的分位数间距差 $\hat{Q}^{(0.75)} - \hat{Q}^{(0.25)}$ 而获得。我们看到教育年限从 11 年变成 12 年,四分位间距增加了 $34426 - 26426 = 8000$。对其他分位数间距使用同样的方法,我们发现,对于样本中间的 80%范围而言,其尺度增加了 15422 美元,而对于中间的 90%范围,则有 19736 美元;对于中间的 95%范围而言,为 28052 美元。

表 5.4 收入分布的尺度变化:从 11 年到 12 年教育年限

分位数与分位数范围	基于样本			基于模型
	教育年限=11 年 (1)	教育年限=12 年 (2)	差异 (2)-(1)	
$Q_{0.025}$	3387	5229	1842	665
$Q_{0.05}$	5352	7195	1843	1130
$Q_{0.10}$	6792	10460	3668	1782
$Q_{0.25}$	12098	18694	6596	3172
$Q_{0.75}$	38524	53120	14596	6598
$Q_{0.90}$	58332	77422	19090	8279
$Q_{0.95}$	74225	95804	21579	9575
$Q_{0.975}$	87996	117890	29894	11567

（续表）

分位数与分位数范围	基于样本			基于模型
	教育年限=11年 (1)	教育年限=12年 (2)	差异 (2)−(1)	
$Q_{0.75}-Q_{0.25}$	26426	34426	8000	
$\hat{\beta}^{*}_{0.75}-\hat{\beta}^{*}_{0.25}$				3426
$Q_{0.90}-Q_{0.10}$	51540	66962	15422	
$\hat{\beta}^{*}_{0.90}-\hat{\beta}^{*}_{0.10}$				6497
$Q_{0.95}-Q_{0.05}$	68873	88609	19736	
$\hat{\beta}^{*}_{0.95}-\hat{\beta}^{*}_{0.05}$				8445
$Q_{0.975}-Q_{0.025}$	84609	112661	28052	
$\hat{\beta}^{*}_{0.975}-\hat{\beta}^{*}_{0.025}$				10902

QRM 拟合为我们提供了估计尺度变化效应的替代方法。这里，我们用符号$\hat{\beta}^{(p)}$表示在一个第 p 分位数回归模型中某个协变量的拟合系数。这一系数表示协变量的单位增加所导致的任何特定分位数下的数值的增加或减少。这样，当协变量增加一个单位时，相应的第 p 分位数间距将改变 $\hat{\beta}^{(1-p)}-\hat{\beta}^{(p)}$ 的量，这里第 p 尺度变化效应用 $SCS^{(p)}$ 表示。

$$
\begin{aligned}
SCS^{(p)} &= IQR_C^{(p)} - IQR_R^{(p)} = (Q_C^{(1-p)} - Q_C^{(p)}) - (Q_R^{(1-p)} - Q_R^{(p)}) \\
&= (Q_C^{(1-p)} - Q_R^{(1-p)}) - (Q_C^{(p)} - Q_R^{(p)}) \\
&= \hat{\beta}^{(1-p)} - \hat{\beta}^{(p)} \quad (\text{当 } p < 0.5)
\end{aligned}
\qquad [5.1]
$$

如果我们拟合一个没有协变量交互项的线性 QRM，那么尺度效应便不会依赖于特定的协变量设置（参照组）。当 $SCS^{(p)}$ 等于 0 时，便不会出现尺度变化。而负值则表明增加协变量数值时会减小尺度，正值则指示相反的效应。

使用方程 5.1 和表 5.2 的估计值，随着教育年限增加 1 年，总体中间 50% 部分的尺度变化是 3426 美元（将第 0.75 分位数的系数减去第 0.25 分位数的系数：6598 − 3172 = 3426）。为什么这一尺度变化会小于观察到的尺度变化（8000 美元）呢？这里有两方面的原因。基于模型的测量是控制了其他协变量

（这里指种族）的部分测量。另外，基于样本分位数的尺度变化，是从两个特定教育组计算得到的，而基于模型的测量则考虑了所有教育组的情况。通过方程5.1，我们可以将教育的QRM系数解释为总体的中间80%部分的尺度增加6497美元，中间90%部分则增大8445美元，而中间95%部分的增长幅度为10902美元（见表5.4最后一列）。

我们可以用相同的方式从尺度变化上解释种族效应。在表5.2中，控制了教育后，白人的收入离散幅度比黑人的更大：对总体中间50%部分而言，为 $12142-6724=5418$，中间80%部分为 $14049-4689=9360$，而中间90%部分则为 $17484-3197=14287$。

当保持原来的偏态不变时，尺度的变化可能成比例地伸展或者缩小位于中位数之上或之下的分布部分。而当原来的偏态发生变化时，它同样可以不成比例地伸展或者缩小位于中位数之上或之下的分布部分。而方程5.1无法区别成比例和不成比例的尺度变化。

偏态变化

一个与更大偏态相关的不成比例的尺度变化，表明存在对因变量分布形状的附加效应。第2章提出了直接测量分位差偏态（quantile-based skewness）的方法，即QSK，定义为上位离散幅度与下位离散幅度的比率的值减去1（见方程2.2）。如果QSK大于0，分布为右偏，反之亦然。图3.2中关于教育组和种族组的箱图表明上位离散情况与下位离散情况的不对称。表5.5的中间部分（分位数排列）的第1和第2列分别描述了11年和12年教育年限组的上位和下位离散情况。我们可以看到：两个教育年限组在样本中间的50%、80%、90%和95%部分都存在右偏的收入分布。

当我们检验比较组的偏态是否不同于参照组的偏态时，我们需要寻找不成比例的尺度变化。图5.4描述了假设情境下关于右偏分布的一种不成比例的尺度变化。让 M_R 和 M_C 分别表示参照组和比较组的中位数。参照组的上位离散为 U_R-M_R，而比较组的则为 U_C-M_C。参照组的下位离散为 M_R-L_R，而比较组的则为 M_C-L_C。不对称性可通过 $(U_C-M_C)/(U_R-M_R)$ 和 $(M_C-L_C)/(M_R$

$-L_R$) 的比率得到。如果这一比率的比值等于 1,即不存在偏态变化。如果它小于 1,那么右偏的幅度在减小。如果它大于 1,则意味着右偏的幅度在增大。以百分率表示的变化可通过将这一比率减去 1 而获得,我们称之为偏态变化 (skewness shift),或者 SKS。

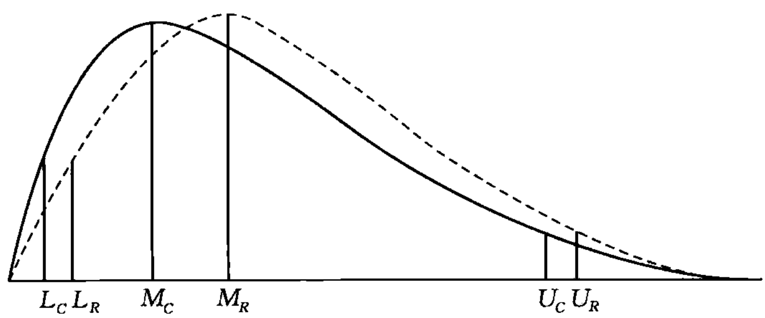


图 5.4 偏态的转换

下面看一下表 5.5 中的样本 SKS——11 年教育年限组和 12 年教育年限组的偏态变化。尽管我们从上一部分内容了解到高教育年限组的尺度大于低教育年限组,但是高教育年限组的右偏程度相对较低,因为样本中间 50%部分的 SKS 是−0.282,中间 80%部分为−0.248,中间 90%部分为−0.283,中间 95%部分为−0.195。因此,各种分位数范围的偏态程度的降低幅度在−19.5%和−28.3%之间。

表 5.5 收入的偏态变化:增加一年教育年限的分布情况

分位数	基于样本			基本模型		
	分位数 (ED = 11)	分位数 (ED = 12)	$SKS^{(p)}$	QRM $\hat{\beta}$	QRM $\hat{\alpha}$	$SKS^{(p)}$
0.025	3387	5229	−0.195	665	6900	−0.049
0.05	5352	7195	−0.283	1130	9850	−0.047
0.10	6792	10460	−0.248	1782	14168	−0.037
0.25	12098	18694	−0.282	3172	24932	−0.016
0.50	20985	32943		4794	42176	
0.75	38524	53120		6598	65745	
0.90	58332	77422		8279	94496	
0.95	74225	95804		9575	120104	
0.975	87996	117890		11567	150463	

（续表）

分位数	基于样本			基本模型		
	分位数 ($ED=11$)	分位数 ($ED=12$)	$SKS^{(p)}$	QRM $\hat{\beta}$	QRM $\hat{\alpha}$	$SKS^{(p)}$
分位数范围						
$Q_{0.75}-Q_{0.50}$	17539	20177				
$Q_{0.50}-Q_{0.25}$	8887	14249				
$Q_{0.90}-Q_{0.50}$	37347	44479				
$Q_{0.50}-Q_{0.10}$	14193	22483				
$Q_{0.95}-Q_{0.50}$	53240	62861				
$Q_{0.50}-Q_{0.05}$	15633	25748				
$Q_{0.975}-Q_{0.50}$	67011	84947				
$Q_{0.50}-Q_{0.025}$	17598	27714				

注：基于样本的 $SKS^{(p)}=[(Q_C^{(1-p)}-Q_C^{(0.5)})/(Q_R^{(1-p)}-Q_R^{(0.5)})]/[(Q_C^{(0.5)}-Q_C^{(p)})/(Q_R^{(0.5)}-Q_R^{(p)})]-1$

对于总体的中间 50%部分，我们有：

$$
\begin{aligned}
SKS^{(0.25)} &= [(Q_C^{(0.75)}-Q_C^{(0.5)})/(Q_R^{(0.75)}-Q_R^{(0.5)})]/[(Q_C^{(0.5)}-Q_C^{(0.25)})/(Q_R^{(0.5)}-Q_R^{(0.25)})]-1 \\
&= [20177/17539]/[14249/8887] \\
&= [1.150/1.603]-1 \\
&= -0.283
\end{aligned}
$$

基于模型的偏态变化是：

$$
\begin{aligned}
SKS^{(0.25)} = {} & [(\hat{\beta}^{(1-p)}+\hat{\alpha}^{(1-p)}-\hat{\beta}^{(0.5)}-\hat{\alpha}^{(0.5)})/(\hat{\alpha}^{(1-p)}-\hat{\alpha}^{(0.5)})]/ \\
& [(\hat{\beta}^{(0.5)}+\hat{\alpha}^{(0.5)}-\hat{\beta}^{(p)}-\hat{\alpha}^{(p)})/(\hat{\alpha}^{(0.5)}-\hat{\alpha}^{(p)})]-1
\end{aligned}
$$

对于总体的中间 50%部分，我们有：

$$
\begin{aligned}
SKS^{(0.25)} &= [(\hat{\beta}^{(0.75)}+\hat{\alpha}^{(0.75)}-\hat{\beta}^{(0.5)}-\hat{\alpha}^{(0.5)})/(\hat{\alpha}^{(0.75)}-\hat{\alpha}^{(0.5)})]/[(\hat{\beta}^{(0.5)}+\hat{\alpha}^{(0.5)}- \\
&\quad \hat{\beta}^{(0.25)}-\hat{\alpha}^{(0.25)})/(\hat{\alpha}^{(0.5)}-\hat{\alpha}^{(0.25)})]-1 \\
&= [(6598+65745-4794-42176)/(65745-42176)]/[(4794+42176- \\
&\quad 3172-24932)/(42176-24932)]-1 \\
&= [25373/23569]/[18886/17244]-1 \\
&= [1.0771/1.094]-1 \\
&= -0.016
\end{aligned}
$$

我们的任务是利用 QRM 系数获得基于模型的 SKS，它涉及参照组的条件分位数。我们将有代表性的协变量组确定为参照类（估计常数 $\hat{\alpha}$）。总体中间的 $100(1-2p)\%$ 部分的 SKS 为：

$$SKS^{(p)}=\frac{\left[\dfrac{Q_C^{(1-p)}-Q_C^{(0.5)}}{Q_R^{(1-p)}-Q_R^{(0.5)}}\right]}{\left[\dfrac{Q_C^{(0.5)}-Q_C^{(p)}}{Q_R^{(0.5)}-Q_R^{(p)}}\right]}-1=\frac{\left[\dfrac{\hat{\beta}^{(1-p)}+\hat{\alpha}^{(1-p)}-\hat{\beta}^{(0.5)}-\hat{\alpha}^{(0.5)}}{\hat{\alpha}^{(1-p)}-\hat{\alpha}^{(0.5)}}\right]}{\left[\dfrac{\hat{\beta}^{(0.5)}+\hat{\alpha}^{(0.5)}-\hat{\beta}^{(p)}-\hat{\alpha}^{(p)}}{\hat{\alpha}^{(0.5)}-\hat{\alpha}^{(p)}}\right]}-1 \quad [5.2]$$

需要注意的是，因为我们求两个比率的比值，所以 SKS 有效地排除了成比例尺度变化的影响。当 $SKS=0$ 时，表明不存在尺度变化或者成比例的尺度变化。这样，SKS 可以测量在成比例尺度变化之上和之外的偏态程度。$SKS<0$ 表示由于解释变量的影响而导致右偏程度的减小，$SKS>0$ 则表明右偏程度的增大。

表 5.5 的右边部分（模型部分）展示了教育的估计系数（$\hat{\beta}$），典型协变量组的估计常数（$\hat{\alpha}$）和基于模型的 SKS。从所有 4 个被选的 SKS 来看，教育年限增加 1 年，会轻微减小右偏程度。变化百分比的减小范围在−1.6%和−4.9%之间。这些基于模型的估计值远小于基于样本的 SKS，因为基于模型的教育净效应是来自典型协变量设定的一种变化，即控制了种族效应。

表 5.6　从黑人到白人收入分布的偏态变化（基于模型）

P	$QRM\hat{\beta}$	$QRM\hat{\alpha}$	SKS^p
0.025	2576	6900	−0.076
0.05	3197	9850	−0.087
0.10	4689	14168	−0.085
0.25	6724	24932	−0.066
0.50	9792	42761	
0.75	12142	65745	
0.90	14049	94496	
0.95	17484	120104	
0.975	22092	150463	

对于白人组而言，其条件收入的偏态较小（见表 5.6）：总体的中间 50%部分的 SKS 为−6.6%，中间 80%部分为−8.5%，中间 90%部分为−8.7%，而中间 95%部分则为−7.6%。它表示相比于总体的中间 50%部分，中间 80%和 90%

部分的右偏程度的减小幅度更大。这一发现表明:白人的中上阶层的收入离散程度要大于黑人的中上阶层。

至此,我们已经提出了协变量对因变量不平等程度的影响效应的总体评估方法,当位置、尺度和偏态的变化在统计上显著时,这一方法可检验这些变化的符号的组合情况。一个正向而显著的位置变化表示比较组的中位数高于参照组的中位数。一个正向而显著的尺度变化则表示比较组的离散程度大于参照组的离散程度。并且,一个正向而显著的偏态变化表示比较组的右偏程度比参照组的大。

如果我们将参照组编码为比较组,将比较组编码为参照组,我们会得到三个负的变化值。因此,这些变化的符号组合,我们称之为"同步"(in-sync)变化,使得总体的收入分布更加不平等和底层的收入水平更加集中。当由一个预测变量产生的这三种变化是同步的时候,这一预测变量经由位置和形状变化等两方面加剧了不平等的程度。如果这些变化的符号是不一致的,则说明预测变量对因变量的位置和形状的改变是朝着相反的方向进行的,并且预测变量对因变量不平等程度的总效应被削弱了。我们将这种情况称为"不同步"模式(out of sync)。

表 5.7 总结了我们收入例子的总体估计情况,还包括自举置信区间。如果这一置信区间在 95%显著水平下包含了 0 值,那么我们就不能确定变化是正向的还是负向的。在表 5.7 中,只有一个变化统计值是不显著的(对于总体的中间 50%部分的白人 SKS)。

表 5.7 形状变化的点估计和 95%置信区间:500 次再抽样自举法

变量	位置 (0.05)	SCS (0.025—0.975)	SKS (0.025—0.975)	SKS (0.05—0.95)	SKS (0.10—0.90)	SKS (0.25—0.75)
收入						
ED	4794	10920	−0.049	−0.046	−0.037	−0.017
下限	4592	10162	−0.056	−0.053	−0.044	−0.028
上限	4966	11794	−0.041	−0.038	−0.029	−0.005
WHITE	9792	19027	−0.079	−0.090	−0.088	−0.067
上限	9474	10602	−0.151	−0.147	−0.152	−0.136
下限	10110	26712	−0.023	−0.037	−0.024	0.005

从表 5.7 可以看到,教育年限增加 1 年,导致了正向的位置和尺度变化,以及负向的偏态变化。这一模式是不同步的。相似的,作为白人则导致正向的位置和尺度变化,以及负向的偏态变化,同样表现为不同步模式。因此,这一简单的模型告诉我们,高教育年限和作为白人是与更高的收入中位数和更大的收入离散范围相关联的,而低教育年限和作为黑人的收入分布的偏态程度更大。如果这一简单模型是正确的,则意味着教育和种族差异都不会加剧收入不平等的程度。当通过"同步"或"不同步"的效应来概括 QRM 的估计情况时,这一例子说明了将变量进行分类的价值所在。一旦我们确定了涉及同步性的变量效应,如上述教育和种族变量时,我们就可以相当容易地判断这一变量是否对分布的不平等程度有所贡献。

小结

本章提出各种方法解释分位数回归模型(QRM)的估计值。除了检验协变量对特定条件分位数的效应外,例如中位数或更低或更高位置的分位数,我们还扩展到关于分布的解释。我们通过例子讲解了 QRM 估计值的图像解释和用 QRM 估计值测量形状变化的定量方法,包括位置变化、尺度变化和偏态变化。家庭收入的例子说明我们可以直接利用 QRM 的估计值来分析协变量对收入不平等的效应。

这一章集中关注对初始单位因变量下的 QRM 的解释。这些解释方法可以直接应用到线性转换的因变量上。然而,为了得到更好的模型拟合,偏态的因变量经常需要进行单调转换。例如,对数转换是右偏分布最常用的一种转换方式。根据因变量表现形式的不同——是初始尺度还是对数尺度,对效应估计值的解释也会有差异。另外,因为一个模型的分析方法用在另一个模型上可能是无效的,所以对建模方式的选择十分重要。由于这一原因,我们将在第 6 章集中讨论由因变量的单调转换引出的具体话题。

第 6 章 | 单调转换 QRM 的解释

当拟合回归模型时，我们常常对右偏的因变量进行对数转换，以保证模型假设在最低程度上被满足。对数转换是实用的，因为它允许以相对方式对预测变量效应进行解释。相似的，对左偏的因变量取平方形式(或者其他大于 1 的乘方)可使新的分布更加对称。这些非直线的单调转换虽然可以改善模型的拟合程度，但不可以维持原来的分布形状。当模型化与特定协变量的变化量相关联的位置和形状变化时，在初始尺度上分析这些变化比在单调转换尺度上更有意义。因此，为了实现对单调转换后的因变量的 QRM 进行更具实质意义的解释，我们需要从转换尺度系数上获得协变量在初始尺度上的效应。本章将通过对数转换的例子讨论两种可以实现这个目标的方法。这些方法可以应用于因变量的任何单调转换形式。

对数尺度上的位置变化

我们首先从位置变化开始。模型化因变量中心位置的一种方法是考虑涉及教育和收入对数的条件均值模型。表 6.1 表明教育水平每增加一年将使条件均值收入提高到 $e^{0.128} = 1.137$ 个单位，即增加 13.7%。表 6.2 中(中间列为 $p = 0.5$) 与之相应的中位数拟合模型的系数则为 0.131，这表明教育增加一年将使收入的条件中位数增加到 $e^{0.131} = 1.140$ 个单位，或者增加 14%。按相对值计算，教育效应对条件中位数的影响更强，而以绝对值衡量，教育效应对条件均值的影响更强，正如第 5 章所示的。

因为增加一个百分比的概念需要详细说明参照组，所以当预测变量是类型

表 6.1　对数收入的经典回归估计:教育和种族效应

变　量	系　数
ED	0.128**
	(0.0020)
WHITE	0.332**
	(0.0160)
Constant	10.497**
	(0.0050)

注:括号内为渐近标准误** $p<0.01$。

变量时,即代表组员身份时,我们要谨慎选择参照类型以方便解释结果。例如,假设我们拟合了一个模型,将收入对数表示为关于种族(黑人/白人)的函数,用0表示黑人,1表示白人。我们的拟合模型LRM(表6.1)表明,系数0.332表示白人的收入高于黑人的量为$e^{0.332}=1.393$,即收入增加了39.3%。另一方面,如果我们采用相反的编码,用0表示白人,1表示黑人,LRM的线性同变性质告诉我们,黑人的系数应该为−0.332。这里,对黑人负向系数的解释并不等同于收入减少39.3%。相反,这个量会是$e^{-0.332}=0.717$,即收入减少28.3%。这一点在更大数值的系数上将表现得更加明显。例如,在第一个模型中的系数2,表示白人的收入将比黑人增加639%,而在第二个模型中,系数则为−2,意味着黑人的收入比白人减少86.5%。我们必须记住的是,当因变量进行对数转换后,改变虚拟变量的参照组会产生两种不同的结果:系数改变了符号,而百分比的改变则转化为倒数$\left(\frac{1}{e^2}=\frac{1}{7.389}=0.135 \text{ 和 } 0.135-1=0.865\right)$。

从对数单位回到初始单位

因变量的对数转换提供了一种分析技术和方法,从而获得对数据更好的拟合效应和以相对形式解释估计值。

在初始单位上的乘法计算成为对数单位上的加法计算。然而,对数转换后的因变量的线性函数明确说明误差项为可加的而不是可乘的,因此改变了原先

表 6.2 对数收入的分位数回归估计：教育和种族效应

	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
ED	0.116**	0.131**	0.139**	0.139**	0.140**	0.140**	0.137**	0.136**	0.134**	0.131**
	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
WHITE	0.429**	0.442**	0.413**	0.399**	0.376**	0.349**	0.346**	0.347**	0.333**	0.323**
	(0.040)	(0.029)	(0.030)	(0.025)	(0.023)	(0.019)	(0.020)	(0.018)	(0.017)	(0.018)
Constant	9.148**	9.494**	9.722**	9.900**	10.048**	10.172**	10.287**	10.391**	10.486**	10.578**
	(0.014)	(0.010)	(0.010)	(0.009)	(0.008)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)
	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	
ED	0.128**	0.129**	0.125**	0.124**	0.121**	0.117**	0.116**	0.117**	0.123**	
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	
WHITE	0.303**	0.290**	0.295**	0.280**	0.264**	0.239**	0.231**	0.223**	0.222**	
	(0.019)	(0.017)	(0.016)	(0.017)	(0.015)	(0.017)	(0.017)	(0.020)	(0.027)	
Constant	10.671**	10.761**	10.851**	10.939**	11.035**	11.140**	11.255**	11.402**	11.629**	
	(0.007)	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)	(0.007)	(0.070)	

注：括号内为渐近标准误，** $p < 0.01$。

误差项的分布。另外,使用对数转换有着明显的缺点,那就是它大大地扭曲了测量单位。在关于不平等的研究中,对数转换的效果是人为地缩小了不平等的表现,因为它大大缩短了分布的右尾部分。特别是,我们对模型化初始单位下因变量的中心位置的影响效应更感兴趣,而不是在对数转换形式下。

对数转换因变量的位置移动导致了初始单位因变量的分布发生怎样的变化呢?这一答案取决于位置估计值的选择。在条件均值的案例中,对数单位的估计值无法提供关于初始单位下的变化信息,反之亦然。只有线性转换拥有同变性质,这就可以通过一个随机变量的均值来计算其转换后的均值。因为对数转换是非线性的,所以条件均值收入不会是对数收入的条件均值的指数函数,正如我们在第 3 章说明的。实际上,在根据对数收入模型的系数以绝对形式计算协变量效应的过程中,不存在简易或者闭型解的表达式。相比之下,中位数回归模型则适合得多。对因变量进行单调转换时,条件中位数亦随之转换。

一般地说,QRM 的单调同变性质保证了对数转换因变量的条件分位数等同于初始单位因变量的条件分位数的对数形式。当这一单调同变性质在总体水平上成立时,由于对数转换形式的非线性特征,将估计值变回原型将复杂得多。麻烦的是,对于连续型变量,受协变量影响的因变量分位数的变化比率取决于这一协变量的真实数值。在类型变量的情况下,组别身份的变化效应同样取决于协变量的数值。无论在哪种情况下,关于协变量对因变量分位数的影响效应,我们必须给出一个准确的含义。我们叙述了解决这一问题的两种方法。第一种需要利用协变量的特定数值,我们称之为典型设定值(Typical-Setting Effects, TSE)。第二种是均值效应(Mean Effect, ME),即在总体中所有相关个案上,对协变量影响条件分位数的效应取平均数。

典型设定值

我们对协变量在绝对形式上影响因变量的效应感兴趣,而处理的方式是确定协变量在典型设定数值上的效应。一种相对直接的方法,是将这一典型设定值看作协变量均值的矢量。如果因变量的均值被表达为协变量的非线性函数,那么在估计其效应时,这便是一种普遍的做法。[13]

下面，我们以二元协变量的例子解释这一想法。从这里，你会了解到如何处理多元协变量的情况。用 x 表示一个连续型协变量（例如，ED），d 表示一个虚拟变量（例如，白人 WHITE）。在本章节的后半部分，我们确定一个特定值 p。在拟合的第 p 分位数回归模型下，我们有：

$$\hat{Q}^{(p)}(\log y \mid x, d) = \hat{\alpha}^{(p)} + \hat{\beta}_x^{(p)} x + \hat{\beta}_d^{(p)} d \qquad [6.1]$$

但是，常数项 $\hat{\alpha}^{(p)}$ 可被解释为在 $x = 0$ 和 $d = 0$ 的情况下，因变量的第 p 分位数的一个估计值。由于协变量常常是非负的，所以对数值的选择不具有特殊的意义，这使得对 $\hat{\alpha}^{(p)}$ 的解释变得有些无趣。另一方面，如果我们将所有协变量以其均值为中心，然后拟合第 p 分位数回归模型：

$$\hat{Q}^{(p)}(\log y \mid x, d) = \hat{\alpha}^{(p)} + \hat{\beta}_x^{(p)}(x - \bar{x}) + \hat{\beta}_d^{(p)}(d - \bar{d}) \qquad [6.1']$$

这为参数 $\hat{\alpha}^{(p)}$ 提供了有着不同解释的另一个拟合值：对数转换因变量的第 p 分位数在协变量特定数值下的一个估计值。其他拟合系数 $\hat{\beta}_x^{(p)}$ 和 $\hat{\beta}_d^{(p)}$ 在方程 6.1 和方程 6.1′中是相同的。

现在思考一下：当我们修改其中一个协变量模型时，结果会发生什么变化，例如，我们从典型设定中将 x 增加一个单位，并维持其他协变量在它们的均值水平上不变。对数因变量的拟合第 p 分位数等于常数项和协变量 x 系数的和：对于 x，为 $\hat{\alpha} + \hat{\beta}_x$；对于 d，则为 $\hat{\alpha} + \hat{\beta}_d$。

我们希望知道这些改变对初始单位因变量的影响。QRM 的单调同变性质告诉我们，如果我们知道对数单位因变量的分布的分位数，这一分位数的指数形式便是初始单位上的分位数。特别的，在典型设定中（所有协变量取均值），对数单位上的条件分位数的指数转换产生初始单位上的拟合条件分位数：$e^{\hat{\alpha}}$。相似的，在修改的协变量数值下，对数单位拟合的条件分位数的指数转换分别等于 $e^{\hat{\alpha}+\hat{\beta}_x}$ 和 $e^{\hat{\alpha}+\hat{\beta}_d}$。根据协变量的单位变化修正过的条件分位数减去在典型设定下的拟合分位数，将得到该协变量在初始单位上的效应，由协变量均值计算得到：对于 x，为 $e^{\hat{\alpha}+\hat{\beta}_x} - e^{\hat{\alpha}}$；对于 d，则为 $e^{\hat{\alpha}+\hat{\beta}_d} - e^{\hat{\alpha}}$。照这样，我们可获得协变量在因变量的任意条件第 p 分位数的效应。

为了理解协变量对因变量的潜在影响，我们最好将对数单位系数转换回初

始单位系数。如果我们打算使用渐近方法，我们必须采用 delta 方法（delta method），而且如果没有封闭解，求解过程将会十分复杂。使用分析方法来推断这些值是不切实际的。相反，我们采用灵活的自举法（在第 5 章所描述的）来获得这些值的标准误和置信区间。

表 6.3 对数收入 QRM 在典型设定效应和均值效应下的点估计和 95%置信区间（500 次再抽样自举法）

	ED			WHITE		
	效应	CI 下限	CI 上限	效应	CI 下限	CI 上限
典型设定效应						
0.025	660	530	821	4457	3405	6536
0.05	1157	1015	1291	4978	4208	6400
0.10	1866	1747	1977	7417	6062	8533
0.15	2486	2317	2634	8476	7210	9951
0.25	3477	3323	3648	10609	8839	12378
0.50	5519	5314	5722	15051	12823	17075
0.75	7992	7655	8277	18788	15669	21647
0.85	9519	9076	9910	19891	16801	22938
0.90	11108	10593	11676	22733	18468	27444
0.95	14765	13677	15662	28131	21181	34294
0.975	18535	19973	19706	41714	33344	51297
均值效应						
0.025	697	554	887	2719	2243	3424
0.05	1241	1073	1396	3276	2875	3868
0.10	2028	1887	2163	4792	4148	5284
0.15	2717	2514	2903	5613	5007	6282
0.25	3799	3620	4008	7228	6343	8089
0.50	5965	5716	6203	10746	9528	11832
0.75	8524	8114	8865	14141	12162	15858
0.85	10082	9581	10559	15429	13362	17329
0.90	11772	11157	12478	17664	14900	20491
0.95	15754	14476	16810	21875	17207	25839
0.975	19836	18007	21235	31419	26192	37014

表 6.3 的上半部分（TSE）展示了绝对形式下的 ED 和 WHITE 对收入的典型设定效应，由所有协变量的均值计算得到，和通过自举法估计得到的 95%置

信区间。对于中位数而言，当种族在平均值处保持不变时，在总体的教育平均水平之上增加 1 年，将使收入增长 5519 美元。将组别身份由黑人换成白人，保持教育在平均值处不变，将带来 15051 美元的收入增长。教育和种族的典型设定效应在低尾处比在上尾处更弱。这些效应比表 5.1 的结果（初始单位收入的拟合）更大。需要注意的是，收入模型和对数收入模型是从不同的拟合方式而得到的两种不同的模型。同样，典型设定效应的估计是基于协变量的均值，然而从拟合初始单位收入得来的系数适用于协变量的所有设定。

均值效应

典型设定方法简单易行，而且提供了关于协变量单位变化对因变量的影响效应的信息。然而，它仅仅考虑到协变量均值的改变所带来的效应。由于这一效应在协变量数值范围内会发生变化，对特定数值的使用可能会导致事实的歪曲。因此，我们引进另一种可能的方法，即从相反的顺序取平均值：首先计算对于协变量每一个可能的取值，协变量单位变化的效应，然后对数据中所有协变量数值的效应取平均值。当因变量的分位数函数以非线性的形式依赖于协变量时，我们打算使用以上方法，例如，在方程 6.1 和 6.1′中，log(y)被表达为协变量的一个线性函数。相反，如果分位数函数是协变量的一个线性函数，那么这两种取均值的方法会产生相同的结果。

对于一个连续型协变量 x 和任意 p，我们问：如果他/她的 x 增加一个单位，其他协变量保持不变时，一个（随机）个案的第 p 条件分位数会改变多少？然后，我们对参照总体中的所有个体的改变量取均值。继续以二元协变量的模型为例，我们可以确定由于 x 的单位增量而带来的分位数变化为：

$$\Delta Q_x^{(p)} = \hat{Q}^{(p)}(y \mid x+1,\ d) - \hat{Q}^{(p)}(y \mid x,\ d) \qquad [6.2]$$

并且平均的分位数变化等于 x 的单位增量在 p 上的均值效应，由 $ME_x^{(p)}$ 表示：

$$ME_x^{(p)} = \frac{1}{n}\sum_{i=1}^{n}[\hat{Q}^{(p)}(y_i \mid x_i+1,\ d_i) - \hat{Q}^{(p)}(y_i \mid x_i,\ d_i)] \qquad [6.3]$$

在我们的模型，即对数收入关于教育和种族的函数中，教育是一个定距变

量。计算方程 6.3 要求：

(1) 通过 $\hat{Q}^{(p)}(y_i \mid x_i, d_i) = e^{\hat{\alpha}^{(p)} + \hat{\beta}_x^{(p)} x_i + \hat{\beta}_d^{(p)} d_i}$，计算得到每个个案被估计的第 p 条件分位数；

(2) 通过 $\hat{Q}^{(p)}(y_i \mid x_i + 1, d_i) = e^{\hat{\alpha}^{(p)} + \hat{\beta}_x^{(p)} (x_i + 1) + \hat{\beta}_d^{(p)} d_i}$，计算得到相应的第 p 条件分位数，如果他/她的教育年限增加一年；

(3) 求这两项的差；

(4) 对这些差取平均值。

对于一个二分协变量，我们希望知道条件分位数的变化，当个体将他/她的组别身份从 $d = 0$ 换成 $d = 1$ 时，保持其他协变量固定不变。在这种情况下，只有当 $d = 0$ 的子群体是相关的，因为将其他组别包括进来将同时改变其他协变量。这样，对于二分 d 而言，分位数的差异便是：

$$\Delta Q_{d,0,1}^{(p)} = \hat{Q}^{(p)}(y \mid x, 1) - \hat{Q}^{(p)}(y) \qquad [6.4]$$

并且 d 的均值效应，表示为 $ME_{d,0,1}^{(p)}$，等于：

$$ME_{d,0,1}^{(p)} = \frac{1}{n_0} \sum_{i:d_i=0} [\hat{Q}^{(p)}(y_i \mid x_i, 1) - \hat{Q}^{(p)}(y_i \mid x_i, 0)] \qquad [6.5]$$

这里 n_0 表示样本中 $d_i = 0$ 的个案数量。

在我们的例子中，WHITE 是一个虚拟变量。计算将被限制在样本的黑人（$WHITE = 0$）。步骤是：

(1) 通过 $\hat{Q}^{(p)}(y_i \mid x_i, d_i = 0) = e^{\hat{\alpha}^{(p)} + \hat{\beta}_x^{(p)} x_i}$，计算得到每个黑人的第 p 条件分位数；

(2) 通过 $\hat{Q}^{(p)}(y_i \mid x_i, d_i = 1) = e^{\hat{\alpha}^{(p)} + \hat{\beta}_x^{(p)} x_i + \hat{\beta}_x^{(p)}}$，计算得到相应的第 p 条件分位数，如果黑人变成了白人；

(3) 求这两项的差；

(4) 对这些差取平均值。

表 6.3 的下半部分(ME)展示了教育和种族的均值效应及它们的 95%置信区间。ED 和 WHITE 的效应都随着 p 增加。教育效应的大小与典型设定效应相似。然而,WHITE 的均值效应随 p 的变化比典型设定效应要广阔得多。

极微效应

对于上述描述的典型设定效应和均值效应这两种方法,我们通过协变量的单位变化来量化它对因变量的效应。当因变量的分位数函数是协变量的一个非线性函数时,这两种方法都是为应付这种情况而设计的。一般而言,这种计算得到的效应并不与单位的大小成比例。例如,教育的单位可以是半年而不是一整年,并且教育年限增加半年的效应并不等于增加一整年教育效应的一半。另外,一些协变量可以被看作是完全连续的。例如,在健康状况的研究中,我们可以将收入看作一个协变量。

一种替代性方法是考虑协变量影响分位数的极微小的变化比率,那就是,通过导数代替有限的差异。例如,假设我们方程 6.1 的一个模型,给出 $\hat{Q}^{(p)}(y \mid x, d) = e^{\hat{\alpha}^{(p)} + \hat{\beta}_x^{(p)}(x-\bar{x}) + \hat{\beta}_d^{(p)}(d-\bar{d})}$,我们有:

$$\frac{d}{dx}\hat{Q}^{(p)}(y \mid x, d) = \hat{\beta}_x^{(p)} e^{\hat{\alpha}^{(p)} + \hat{\beta}_x^{(p)}(x-\bar{x}) + \hat{\beta}_d^{(p)}(d-\bar{d})}$$

因此,使 $x = \bar{x}$ 和 $d = \bar{d}$,典型设定效应的相似体则变成 $\frac{d}{dx}\hat{Q}^{(p)}(y \mid x, d) = \hat{\beta}_x^{(p)} e^{\hat{\alpha}^{(p)}}$。相似的,均值效应的相似体的形式为:

$$ME_x^{(p)} = \frac{1}{n}\sum_{i=1}^{n}\frac{d}{dx}\hat{Q}^{(p)}(y \mid x_i, d_i) = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_x^{(p)} e^{\hat{\alpha} + \hat{\beta}_x^{(p)}(x_i-\bar{x}) + \hat{\beta}_d^{(p)}(d_i-\bar{d})}$$

对数单位系数的图解

图 6.1 是对数单位系数的图解,分别展示了对数收入 QRM 中教育(ED),白人(WHITE)及其常数的曲线。在图 6.1 中,典型设定下对数收入的条件分

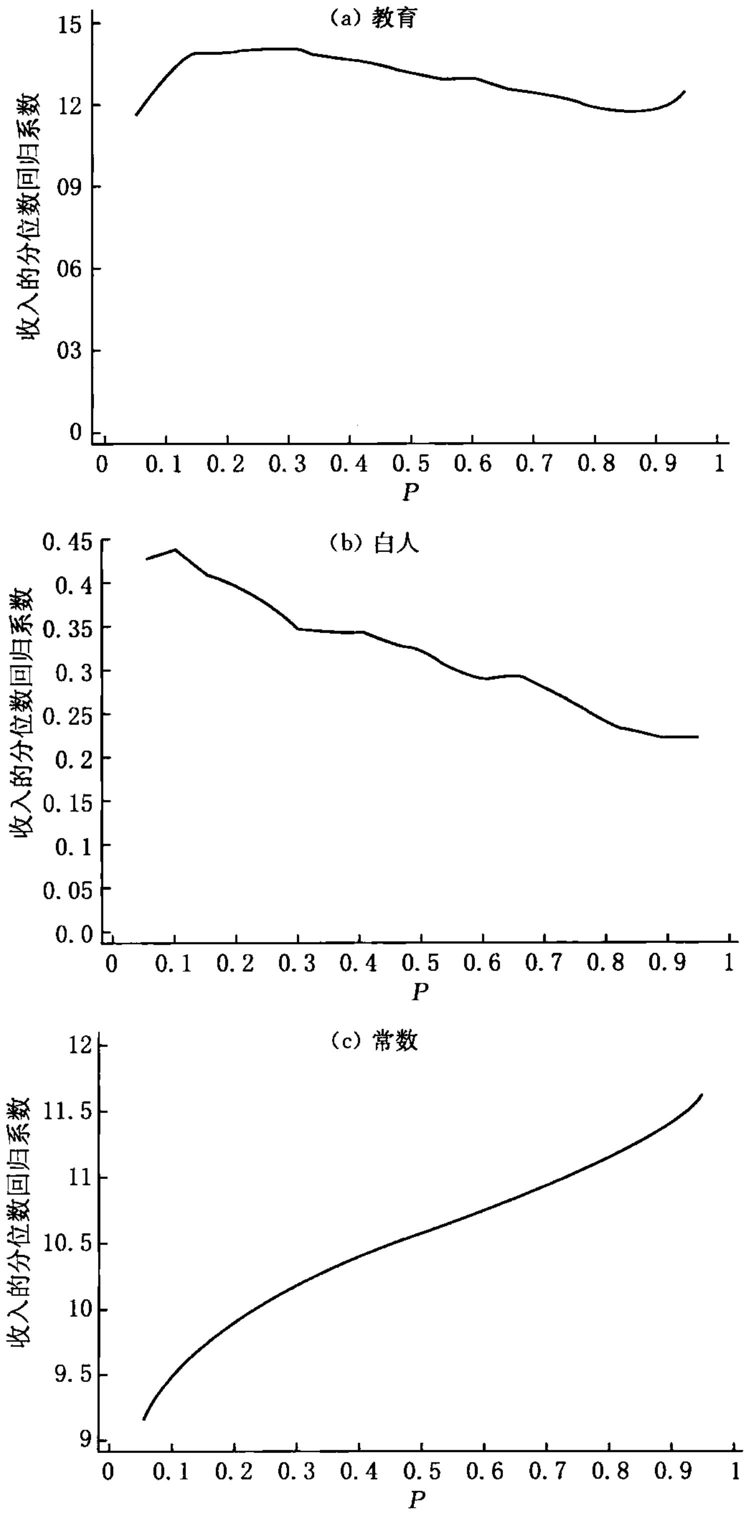


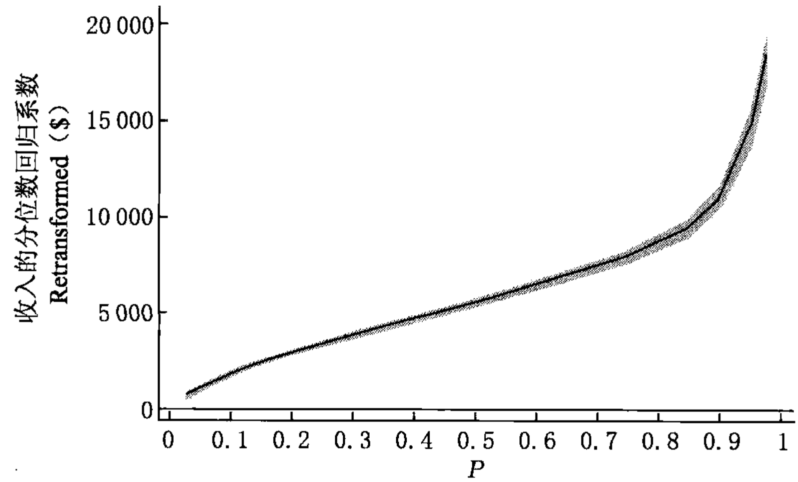
图 6.1　从收入对数 QRM 得到的对数 R 度估计的图形显示

位数函数拥有类似于正态分布的分布形状，因为其中位数以上和以下的斜率是相近的。这一发现表明对收入作对数转换缩短了右尾分布，使得经过转换的分布接近于正态分布。由于对数系数可被理解为百分比的改变，一条水平直线应该表明不发生偏态变化情况下的纯单位转变。任何非水平的曲线要么表示偏态变化，要么表示纯位置变化，但它无法告诉我们确切的是哪一种变化。我们观察到 ED 和 WHITE 曲线是非水平的，所以我们知道它们的效应不纯粹是单位的改变。

然而，我们并不确定这些曲线是表示纯位置的变化还是意味着另外的偏态变化。如果非水平曲线的不确定性是基于对数单位系数的，那么在初始单位上重新计算协变量的效应来分析形状的变化便十分重要。相比之下，基于绝对效应的曲线图可以告诉我们，协变量是否同时导致位置和单位的变化，以及是否会引起偏态变化。例如，采用典型设置效应（TSE），我们可以检视协变量在改变因变量形状方面的作用。

为了同时捕捉位置和单位的变化，图 6.2 展示了对数收入 QRM 中绝对形式下的 ED 和 WHITE 的 TSE 以及它们的置信封闭间。TSE 的图像形状与图 5.1 的十分相似。ED 和 WHITE 都对位置变化、单位变化和可能存在的偏态变化有影响。

(a) 教育



(b) 白人

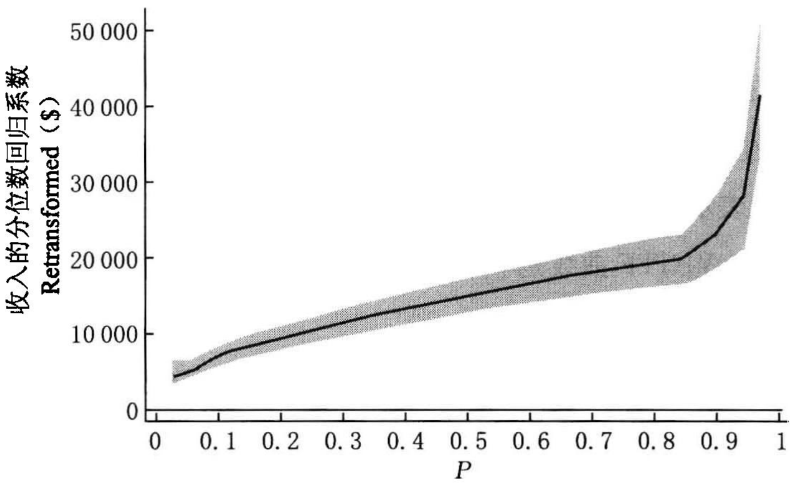


图 6.2 从收入对数 QRM 得到的 TSE(绝对数)的图形显示

从对数单位拟合测量形状变化

因为在初始单位下我们更容易解释形状变化,所以最佳的方法是从对数单位系数中计算初始单位下的形状变化。根据单位变化方程 5.1 和偏态变化方程 5.2,参照组的尺度和偏态情况是进行比较所必须的。当对初始单位的因变量进行拟合时,这些系数就失去了任何参照。然而,当对对数单位的因变量进行拟合时,协变量的数值变化所带来的效应是与不同的参照组相关联的。因此,当对对数单位的因变量进行拟合时,我们需要一个固定的参照组,以便理解形状的变化情况。典型设定效应可以很好地达到这一目标。将表 6.3 的 TSE 结果代入方程 5.1 和方程 5.2 中,我们通过自举再抽样样本计算出尺度变化、偏态变化和它们的置信封闭间,如表 6.4 上半部分所示。ED 和 WHITE 在 $Q_{0.025}$ 到 $Q_{0.975}$ 的范围内都具有正向的尺度变化效应,和在 $Q_{0.25}$ 到 $Q_{0.75}$, $Q_{0.10}$ 到 $Q_{0.90}$, $Q_{0.05}$ 到 $Q_{0.95}$, $Q_{0.025}$ 到 $Q_{0.975}$ 的范围内具有负向的偏态变化效应。而这些测量值的 95%置信区间表明 ED 和 WHITE 的 SKC 是显著的,但是 ED 在四种分位数差中的 SKC 都是显著的,而 WHITE 只有在两种情况下 SKC 才是显著的。由于这些测量值是在协变量的均值处估计得到的,而且对数收入模型不同于收入

模型，所以这些测量值的大小并不等于第5章中提到的那些测量值。但是，我们认为：这些变化值的符号和总体的效应模式应该保持一致。不管是对收入还是对数收入进行拟合，与每一个协变量相联系的位置和形状的变化并不是同步的。

表6.4 对数收入的经典回归估计(教育和种族效应)

变　量	SCS (0.025—0.975)	SKS (0.025—0.975)	SKS (0.05—0.95)	SKS (0.10—0.90)	SKS (0.25—0.75)
典型设定效应					
ED	17861	−0.016	−0.017	−0.025	−0.015
下限	16325	−0.028	−0.029	−0.036	−0.029
下限	19108	−0.006	−0.006	−0.014	0.002
WHITE	37113	−0.010	−0.118	−0.111	−0.090
下限	29014	−0.129	−0.194	−0.193	−0.199
上限	46837	0.054	−0.022	−0.015	0.047
均值效应					
ED	19118	−0.016	−0.014	−0.025	−0.015
下限	17272	−0.028	−0.030	−0.036	−0.029
上限	20592	−0.006	−0.006	−0.014	−0.002
WHITE	28653	−0.046	−0.114	−0.107	−0.084
下限	23501	−0.128	−0.181	−0.175	−0.174
上限	34348	0.042	−0.030	−0.026	0.031

TSE可以用来直接计算协变量对尺度和偏态变化的效应，而不能计算均值效应。但是，协变量对尺度变化和偏态变化效应的导数与均值效应的导数相似。用S表示形状的测量值(尺度或偏态)，ΔS表示形状变化的测量值。对于连续型协变量而言，ΔS的导数是：

$$\Delta S_x^{(p)} = S^{(p)}(y \mid x+1,\ d) - S^{(p)}(y \mid x,\ d) \qquad [6.6]$$

而对于二元协变量，则为：

$$\Delta S_{d,0,1}^{(p)} = S^{(p)}(y,\ d=1) - S^{(p)}(y,\ d=0) \qquad [6.7]$$

对条件分位数的均值效应采取相同的步骤，我们可以从对数收入QRM中

计算尺度变化和偏态变化的均值效应(见表 6.4 的下半部分)。教育年限增加一年导致正向的尺度变化,这与基于 TSE 的效果相似。WHITE 对尺度变化有正向效应,而且它的大小比基于 TSE 的效应要大。而在 ME 和 TSE 之间,教育和种族对偏态变化的效应是非常相似的。ME 下的总体的效应模式也不是同步的,这和 TSE 的情况相同。

小结

本章讨论了 QRM 中由于因变量的非线性单调转换而出现的解释问题。由于 QRM 的单调同变性,我们得以重新计算基于因变量分布的初始单位之上的协变量效应,而这在 LRM 中是不可获得的。虽然如此,这一重新计算需要特殊的方法。本章提出了两种方法。典型设定方法在计算上相对简单,而均值效应方法则稍微复杂一些。两种方法都涉及对协变量数值取平均数,但以不同的规则进行。典型设定效应和均值效应都是在谈论整个样本或子样本。研究者应该选择最适合特定研究问题的方法。

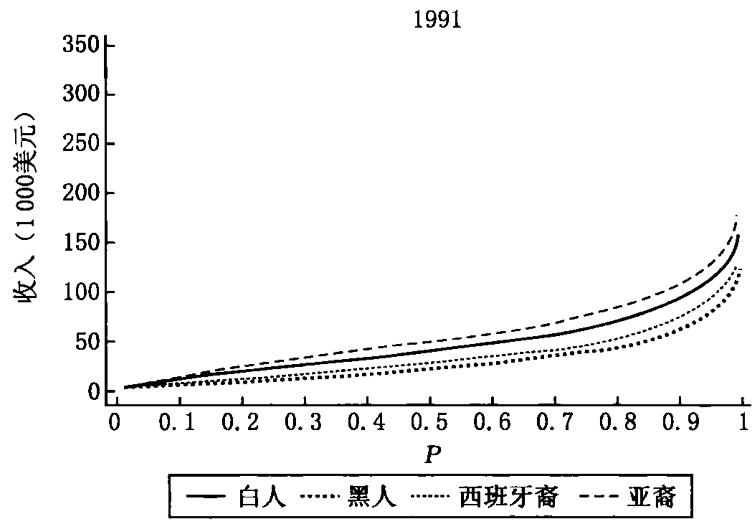
下一章提供了本书所介绍的技术的一个总结,通过将它们应用到真实的研究问题中来阐述。在应用中,我们比较了 1991 年和 2001 年美国收入不平等的来源,讲述了采用 QR 分析的动机和怎么按步骤进行,并给出了全面的 Stata 命令。

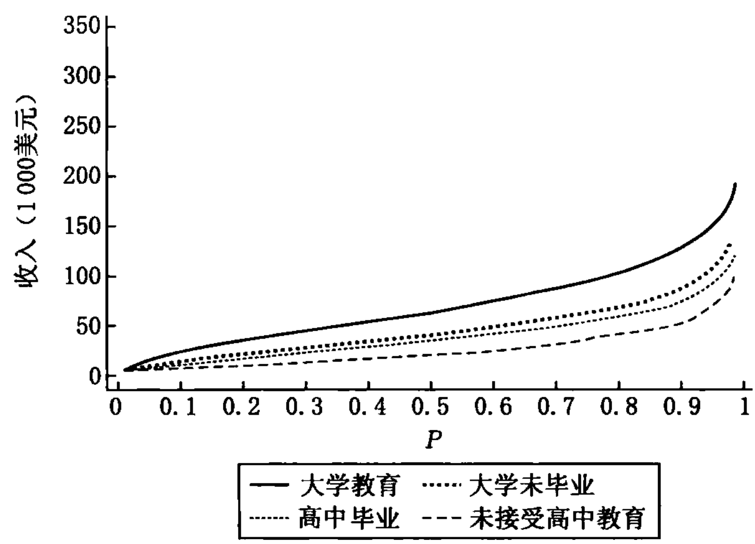
第 7 章 | 实例:1991 年和 2001 年的收入不平等

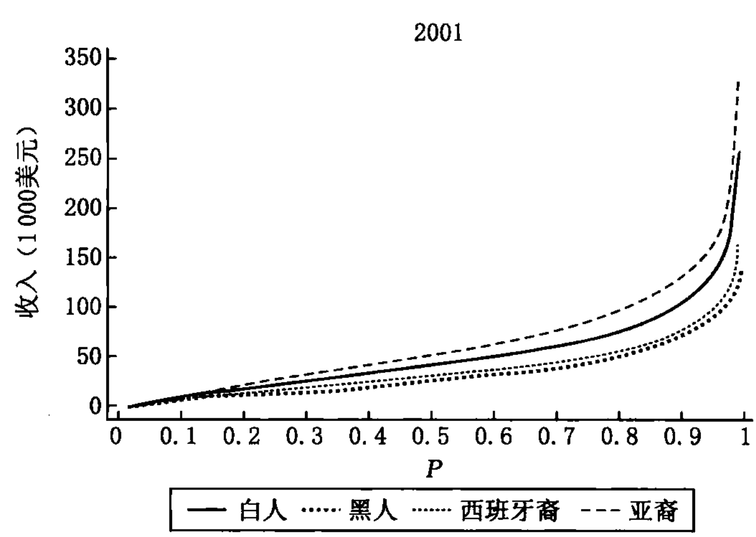
在前面章节中使用的实证说明只限制在一个或两个协变量例子上。本章将本书的技术应用在一个特别议题上:从 1991 年到 2001 年间家庭收入不平等的持续和扩大。我们的目标是通过具体的实证例子系统地概括本书发展出的统计技术。我们从美国“收入和项目参与调查(SIPP)”中提取 1991 年的数据,并将之合并到之前使用的 2001 年数据。家庭收入根据 2001 年固定币值进行了调整。我们明确指定了一个精简模型,即家庭收入是 5 大因素(13 个协变量)的一个函数:生命周期(年龄和年龄的平方),种族/民族(白人、黑人、西班牙人和亚洲人),教育(大学毕业,大学未毕业,高中毕业和非高中学历),家庭类型(有子女的已婚夫妇、没有子女的已婚夫妇、有子女的单亲母亲、单身和其他),农村居民。以上的说明将在本章中用到。并且我们同时拟合初始单位收入模型和对数转换收入模型。分析包括:(a)分别评估初始单位收入和对数单位收入两种模型的拟合优度;(b)比较一般最小二乘法和中位数回归的估计值;(c)对系数进行双尾检验;(d)用图像展示 19 组系数估计值及其置信区间;并且(e)对每一年获得每一个协变量对条件分位数的位置和形状变化的影响效应,并检验年代的发展趋势。

观察到的收入差别

图 7.1 描述了种族/民族组和教育组在 1991 年和 2001 年的 99 个实际的分位数。其中最有趣的特征是与 1991 年相比,2001 年每组中间的 98%成员的收入分散更广。


1991
收入（1 000美元）
350
300
250
200
150
100
50
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
P
白人
黑人
西班牙裔
亚裔


收入（1 000美元）
350
300
250
200
150
100
50
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
P
大学教育
大学未毕业
高中毕业
未接受高中教育


2001
收入（1 000美元）
350
300
250
200
150
100
50
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
P
白人
黑人
西班牙裔
亚裔

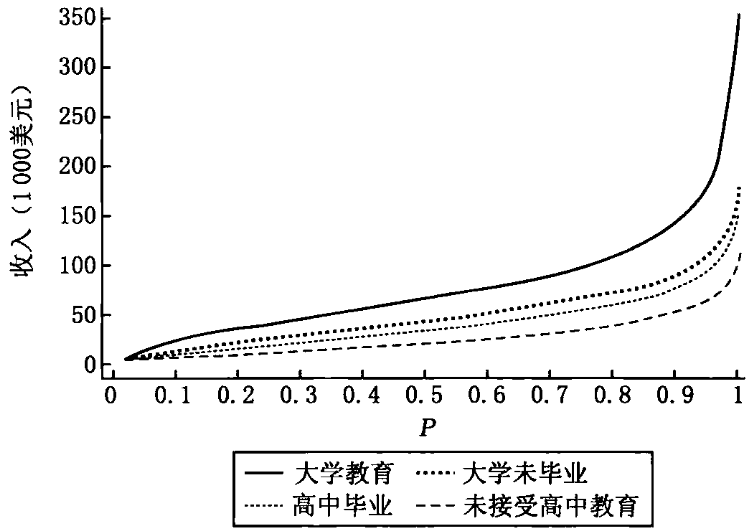


图 7.1 分种族和教育程度的经验分位数函数

更详细的比较需要分位数的确切数值。表 7.1 比较了在 1991 年和 2001 年第 0.025 分位数，中位数和第 0.975 分位数的家庭收入（以 2001 年固定币值计算），并对这些数值赋权，以反映总体情况。观察到的总体和每个组别的共同特征是：相比于 1991 年，2001 年的中间 95%家庭的离散程度（$QSC_{0.025}$）更大，这表明在这 10 年中收入的总体差别和组内差别在增大。

表 7.1 家庭收入的组别分布情况(1991 年和 2001 年)

	分位数					
	1991			2001		
	0.025	0.500	0.975	0.025	0.500	0.975
总体	6256	38324	131352	6000	40212	164323
种族/民族						
白人	6765	40949	135443	6600	42878	172784
黑人	3773	23624	101160	3788	27858	113124
西班牙人	5342	28851	114138	5600	33144	119454
亚洲人	5241	49354	149357	4800	55286	211112
教育						
大学毕业	11196	64688	168912	10910	65298	263796

（续表）

	分位数					
	1991			2001		
	0.025	0.500	0.975	0.025	0.500	0.975
大学未毕业	8059	42082	120316	6364	41901	134796
高中毕业	6392	35723	104102	5347	33246	118162
高中以下	4918	20827	80603	4408	20319	79515
家庭类型						
已婚有子女	12896	55653	143343	14193	61636	204608
已婚无子女	11621	43473	146580	10860	47665	176375
单亲母亲	3666	23420	94114	3653	27690	96650
单身	4884	20906	83213	3977	21369	91551
其他类型	7301	37896	115069	6600	41580	150123
居住地						
城市	6330	40732	137574	6199	42504	174733
农村	6122	32874	111891	5419	33505	118079

在过去的 10 年里，白人和其他种族的差别在收入分布的下半部分缩小了。这一缩小可以被看作白人家庭的第 0.025 分位数收入的减少，相比之下，黑人和西班牙人的收入相应得到适度的增长。在中位数和第 0.975 分位数收入处，亚洲人的收入比白人增长得更快，但底端 2.5%的亚洲家庭收入落后于白人家庭。

收入不平等的一个重要变化是教育对顶端部分的回报。当大多数大学毕业生在过去 10 年里获得了丰厚的收入时，一半以上的非大学毕业生觉得他们的收入实际上在下降。特别的，相对于 1991 年的高中辍学者，2001 年，超过 97.5%的高中辍学者的收入显著下降了。

对家庭类型的考虑——由婚姻状况和子女数量来定义——让我们转向另一个社会分层重新塑造收入分布的领域。有子女的已婚夫妇收入在增加，而单亲母亲家庭和单身家庭的收入变化不大。城市和农村之间的不平等及它们内部的不平等在这 10 年里得到了强化。

描述统计值

表 7.2 展示了分析中各个协变量的加权平均数和标准差。我们发现:从 1991 年到 2001 年间的平均收入几乎增长了 5000 美元,这一增长幅度大于之前表格中观察到的中位数收入的增长。对数收入的小幅增长提醒我们:对数转换缩短了分布的右尾。我们发现:种族/民族结构有着更大的差异,而总体教育水平则出现了大幅提高。然而,有子女的已婚夫妇家庭数量减少了,而其他类型家庭和单身家庭的数量相应增加了。正如在过去 10 年中所见的,美国继续着城市化和市郊化的进程。

表 7.2 变量的描述性统计值

变 量	1991		2001	
	均 值	标准差	均 值	标准差
因变量				
收入($)	46168	33858	51460	46111
对数收入	10.451	0.843	10.506	0.909
年龄	49	17	49	17
年龄平方	2652	1798	2700	1786
协变量				
种族/民族				
白人	0.795	0.404	0.755	0.430
黑人	0.101	0.301	0.094	0.292
西班牙人	0.079	0.269	0.094	0.292
亚洲人	0.025	0.157	0.033	0.177
教育				
大学毕业	0.230	0.421	0.261	0.439
大学未毕业	0.210	0.407	0.296	0.457
高中毕业	0.341	0.474	0.302	0.459
高中以下	0.219	0.414	0.141	0.348
家庭类型				
已婚有子女	0.330	0.470	0.287	0.452
已婚无子女	0.224	0.417	0.233	0.423

（续表）

变 量	1991		2001	
	均 值	标准差	均 值	标准差
单亲母亲	0.108	0.310	0.104	0.305
单身	0.257	0.437	0.267	0.442
其他类型	0.082	0.274	0.110	0.313
居住地				
城市	0.732	0.443	0.773	0.419
农村	0.268	0.443	0.227	0.419

收入调查数据记录

收入调查数据的两个特征，使得 QRM 方法在分析上优于 LRM 方法，只有 0.2%家庭的收入超过 100 万美元，而超过总体 96%的家庭收入低于 10 万美元。因此，特别富有的家庭的数据会严重影响 OLS 的系数估计。第二，收入调查对每项收入资料通常进行顶端编码(top-coded)；因此，我们不能直接确定家庭总收入在哪个水平上被删截。另外，不同年份的调查可能使用不同的顶端编码标准，这导致在清理不同年份的数据以用于比较时十分麻烦。分位数回归模型并不需要考虑这些问题，因为 QRM 具有在第 3 章描述的稳健特性。在这个例子中，我们选择两个极端点，第 0.025 和第 0.975 分位数，这样集中于对总体中间 95%个案进行建模。由于采用顶端编码的数据点对于拟合的第 0.975QRM 而言，倾向于出现正残差，而替代通过顶端编码得到的那些(未知)收入数值对 QRM 估计值的影响效应倾向于最小值。这简化了数据处理，因为我们可以分析调查到的所有数据点，不管是否进行顶端编码。

纵观这个例子，每个协变量向其中位数集中。因此，收入 OLS 回归中的常数项表示总体收入的均值，而对数收入 OLS 回归的常数项则表示对数收入的均值。对建立在中心化协变量之上的 QRM 拟合模型而言，收入分位数回归的常数项表示在典型设定效应下的收入的条件分位数，而对数收入分位数回归的常数项则表示在典型设定效应下的对数收入的条件分位数。

拟合优度

因为 QRM 不再做出线性回归的假设，所以初始单位收入可以不经过转换而直接使用。然而，如果对数转换可以提供更好的拟合模型，我们也愿意进行转换。因此，我们对收入方程和对数收入方程的拟合优度进行了比较。通过 Stata 的“qreg”命令，我们分别在 19 个等距分位数处拟合不同的 QRM（总共有 $2 \times 19 = 38$ 种拟合情况）。尽管 qreg 命令可以产生渐近的标准误（可能是有偏的），但我们只对拟合优度统计值 QRM Rs 感兴趣。表 7.3 分别展示了初始和对数单位因变量的 QRM Rs（见第 5 章的定义）。

表 7.3　拟合优度：初始单位与对数收入的 QRM

分位数	1991			2001		
	收入	对数收入	差异	收入	对数收入	差异
	(1)	(2)	(2)−(1)	(1)	(2)	(2)−(1)
0.05	0.110	0.218	0.109	0.093	0.194	0.101
0.10	0.155	0.264	0.109	0.130	0.237	0.107
0.15	0.181	0.281	0.099	0.154	0.255	0.101
0.20	0.198	0.286	0.088	0.173	0.265	0.091
0.25	0.212	0.290	0.078	0.188	0.270	0.083
0.30	0.224	0.290	0.067	0.200	0.274	0.074
0.35	0.233	0.290	0.057	0.209	0.275	0.066
0.40	0.242	0.289	0.048	0.218	0.277	0.059
0.45	0.249	0.288	0.039	0.225	0.276	0.051
0.50	0.256	0.286	0.029	0.231	0.275	0.044
0.55	0.264	0.282	0.019	0.236	0.273	0.037
0.60	0.270	0.279	0.009	0.240	0.270	0.030
0.65	0.275	0.275	−0.001	0.243	0.266	0.023
0.70	0.280	0.270	−0.010	0.246	0.262	0.015
0.75	0.285	0.264	−0.021	0.249	0.256	0.008
0.80	0.291	0.258	−0.032	0.249	0.250	0.000
0.85	0.296	0.250	−0.047	0.250	0.242	−0.008
0.90	0.298	0.237	−0.061	0.252	0.233	−0.019
0.95	0.293	0.213	−0.080	0.258	0.222	−0.036

一般而言，相对于初始单位，对数转换会导致对数据更好的模型拟合。从1991年的数据来看，在 $0 < p < 0.65$ 时，对数收入的 R 会更高：将近19个分位数的2/3获得了更优的拟合。对2001年的数据来说，在 $0 < p < 0.85$ 时，对数收入的 R 更高，表明对2001年数据采用对数转换比1991年数据具有更强的说服力。然而，对数单位下对上尾部分的拟合并不好。如果我们主要关心上尾部分的变化和分层，应该使用初始单位收入。出于这个原因，我们将讲述两种单位下的分析情况。

条件均值回归与条件中位数回归

我们对条件中位数进行建模，是为了展示收入的中心位置与协变量的关系。相反，条件均值模型（如OLS）估计的条件均值倾向于捕捉收入分布（右偏）的上尾情况。而中位数回归是通过使用Stata的“qreg”命令而获得的。这一命令同样被用在初始样本的500个自举样本上，以获得自举标准误（见附录中关于计算的Stata命令）。

表7.4列出了2001年初始单位和对数单位收入的OLS估计值和中位数回归估计值。我们期待OLS下的效应会强于中位数回归下的效应，因为上尾收入数据对OLS系数存在影响。

收入方程中的系数是以绝对形式出现的，而对数收入系数则是以相对形式表示。除了一些例外，对数收入下的OLS系数的绝对值比对数收入下的中位数回归系数的绝对值大。例如，在OLS结果中，相对于白人，黑人的条件均值收入下降了 $100(e^{-0.274}-1)=-24\%$；但这一降幅在中位数回归结果中则是 $100(e^{-0.2497}-1)=-22\%$。换言之，当控制了其他效应后，黑人的均值收入比白人的低24%，而黑人的中位数收入比白人的低22%，请注意：我们可以在绝对形式下确定黑人对条件中位数的效应，因为QRM具有单调同变性质；但我们无法通过条件均值的对数单位估计值得到绝对效应，因为LRM并不具备单调同变性质。稍后，我们将转向从对数收入方程估计值中获取绝对形式下的效应。

表 7.4 OLS 和中位数回归(2001 年的初始和对数收入)

变量	OLS		中位数	
	系数	标准误	系数	标准误
收入				
年龄	2191	(84.1)	1491	(51.4)
年龄平方	−22	(0.8)	−15	(0.5)
黑人	−9800	(742.9)	−7515	(420.7)
西班牙人	−9221	(859.3)	−7620	(551.3)
亚洲人	−764	(1369.3)	−3080	(1347.9)
大学未毕业	−24966	(643.7)	−18551	(612.5)
高中毕业	−32281	(647.4)	−24939	(585.6)
高中以下	−38817	(830.0)	−30335	(616.4)
已婚无子女	−11227	(698.5)	−11505	(559.6)
单亲母亲	−28697	(851.1)	−25887	(580.2)
单身	−37780	(684.3)	−32012	(504.8)
其他类型	−14256	(837.3)	−13588	(672.8)
农村	−10391	(560.7)	−6693	(344.1)
常数	50431	(235.2)	43627	(185.5)
对数收入				
年龄	0.0500	(0.0016)	0.0515	(0.0016)
年龄平方	−0.0005	(0.00002)	−0.0005	(0.00001)
黑人	−0.2740	(0.0140)	−0.2497	(0.0145)
西班牙人	−0.1665	(0.0162)	−0.1840	(0.0185)
亚洲人	−0.1371	(0.0258)	−0.0841	(0.0340)
大学未毕业	−0.3744	(0.0121)	−0.3407	(0.0122)
高中毕业	−0.5593	(0.0122)	−0.5244	(0.0123)
高中以下	−0.8283	(0.0156)	−0.8011	(0.0177)
已婚无子女	−0.1859	(0.0132)	−0.1452	(0.0124)
单亲母亲	−0.6579	(0.0160)	−0.6214	(0.0167)
单身	−0.9392	(0.0129)	−0.8462	(0.0136)
其他类型	−0.2631	(0.0158)	−0.2307	(0.0166)
农村	−0.1980	(0.0106)	−0.1944	(0.0100)
常数	10.4807	(0.0044)	10.5441	(0.0045)

收入和对数收入方程中 QRM 估计值的图像化

QRM 与 LRM 重要的不同在于：QRM 估计了多组分位数系数。我们使用 Stata 的“sqreg”命令，同时拟合 QRM 的 19 个等距分位数（第 0.05，…，第 0.95）。“sqreg”命令采用自举方法估计这些系数的标准误。我们确定了 500 个复制样本以保证自举样本足够大，以得到稳定的标准误和 95%置信区间估计值。“sqreg”命令没有从每一次自举中保存估计值，而仅仅展示了结果的摘要。我们对初始单位收入和对数转换收入都进行自举抽样。从“sqreg”得到的结果用于系数的图像展示。

使用如此多的估计值，导致了在复杂性和简约性之间的权衡考虑。一方面，众多的参数估计值可以捕捉到分布形状复杂而细微的变化，而这正是使用 QRM 的一大优势。另一方面，这种复杂并不是没有代价的，因为我们可能面对解释系数估计值集合的复杂问题。因此，之前备选的 QRM 估计值的图像视角，便成为一个解释 QRM 结果的重要步骤。

我们特别感兴趣的是：协变量效应在不同分位数上是如何变化的。我们根据被估计的 QRM 系数随着 p 的变化情况而绘制的图像，对于突出这些系数的趋势是十分有价值的。对于初始单位系数，水平线表明系数并不随着 p 变化，因此，协变量的特定变化对因变量分位数的影响效应对所有分位数而言是相同的。换言之，当所有其他协变量保持不变时，这个协变量的变化只引起位置的变化：如果这条直线在水平零直线之上，那么存在正向的变化；如果在水平零直线之下，那么变化是负向的。另一方面，一条非水平直线代表位置和尺度都发生了变化。在这种情况下，位置的变化由中位数处的分位数系数决定：正的中位数系数表明向右的位置变化，负的中位数系数则表明向左的位置变化。一条向上倾斜的直线表明正向的尺度变化（尺度越来越宽）。相反，向下倾斜的直线表明负向的尺度变化（尺度越来越窄）。曲线中任何非直线的形状意味着存在更复杂的形状变化，例如，以偏态变化的形式。然而，这些图像既不提供确切的形状变化的分位数，也不提供它们的统计显著性。稍后，我们将通过形状变化

分位数检验它们的显著性。

为了说明如何通过图像化识别位置和形状的变化情况，我们在图 7.2 中仔细检验了年龄对初始单位收入的影响效应。因为系数和置信封闭间都大于 0（那条水平线），年龄对初始单位收入的各种分位数的效应全都是正向和显著的。这些年龄系数形成一条向上倾斜的、近似笔直的线，这表明：年龄的增长使得收入分布的位置向右移动，而且扩大了收入分布的尺度。

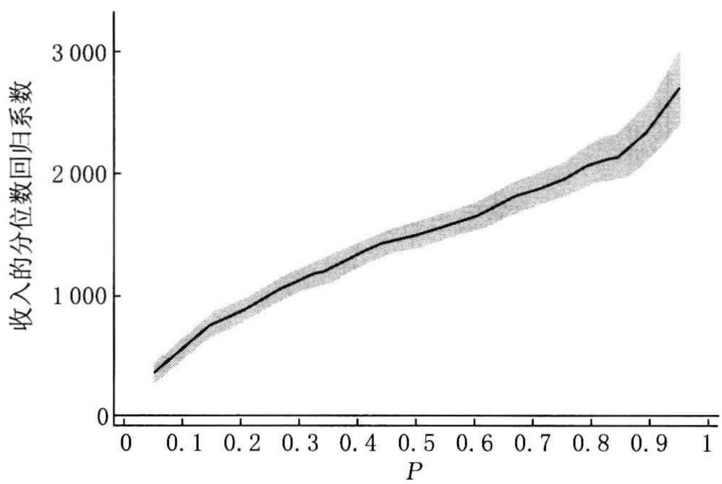


图 7.2　年龄效应：原始尺度的 QRM 系数和 BOOSTRAP 置信区间(2001)

图 7.3 的小图展示了初始单位收入的结果。基于自举标准误计算得到的系数点估计和 95%置信区间在 $p \in (0, 1)$ 范围内被绘制成图像。图中的阴影部分表明：如果它没有穿过零值，那么这个协变量的效应在特定分位数上是显著的。例如，亚洲人效应在 $p > 0.45$ 往后处是不显著的，因为置信封闭间在这点之后穿过了 0 点。第 4 章总结了一些基本模式，它提供了关于初始和对数单位系数的位置变化和尺度变化的一些提示。下面将讨论我们的例子中出现的模式。

常数值的图像是关于特定家庭收入的被预测的分位数函数（例如，基于所有协变量的均值而虚构的家庭收入），它将充当参照函数。这一分位数函数表明：对于特定的家庭而言，其收入是一个右偏分布。与没有考虑协变量的影响而从收入数据观察到的偏态程度相比，这一偏态更不明显。在这 13 个协变量

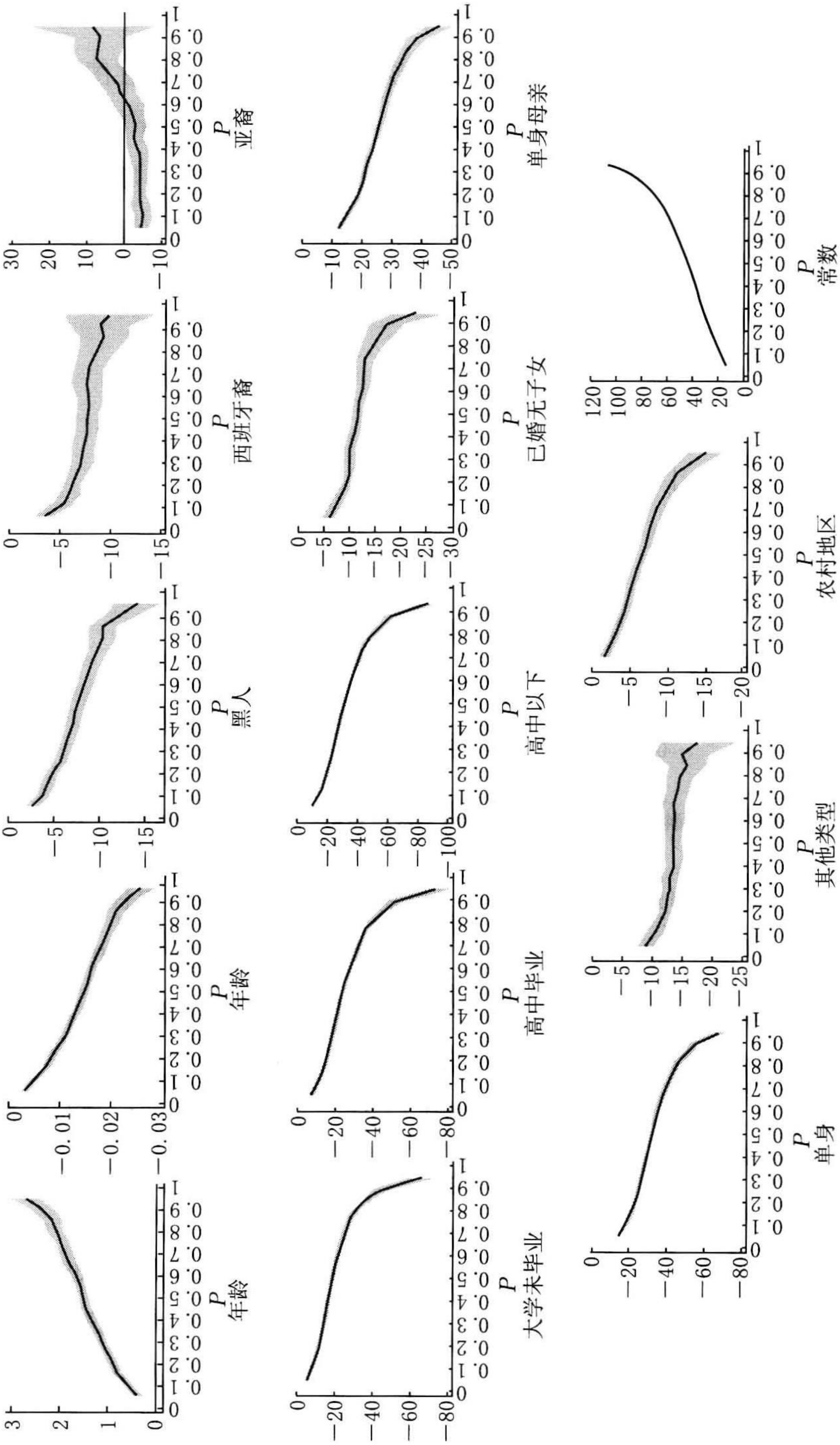


图 7.3 原始尺度的 QRM 系数和 BOOSTRAP 置信区间(2001)

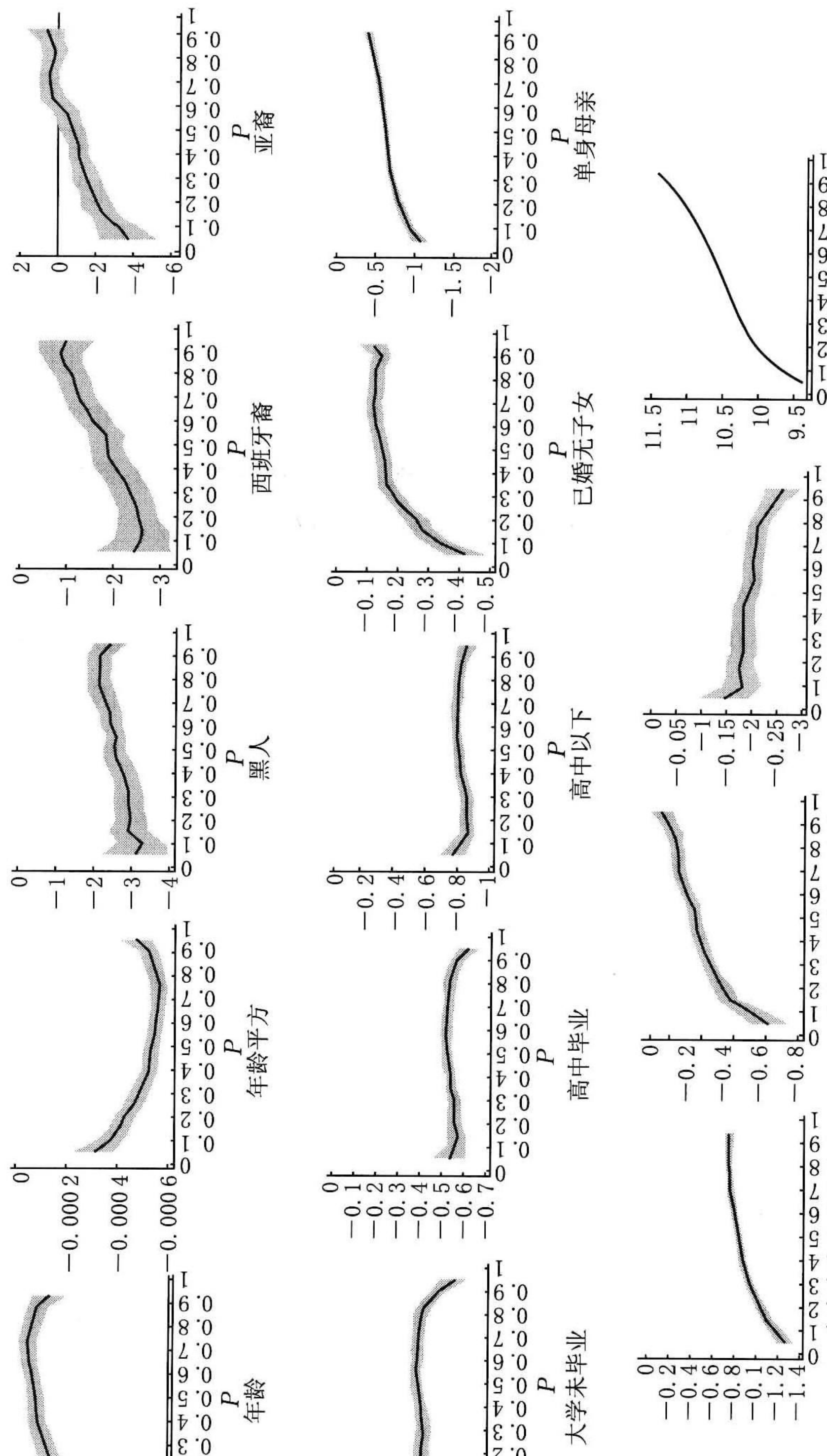


图 7.4 取对数的 QRM 系数和 BOOSTRAP 置信区间(2001)

中，只有"年龄"具有正效应。总体中间的 70%个案的估计收入随着年龄的增长而成比例地增加。年龄效应在低尾部分的比率过低，而上尾部分的比率却过高。然而，这种不对称性并不足以得出关于偏态的结论，因为必须考虑基准的偏态程度（由常数项表示）。所有其他协变量的效应都是负的。正如之前提到的，亚洲人效应在条件分布的低尾部分是显著的。这部分的曲线比较平直，表示分布的下半部分仅存在位置变化。另外一些协变量也存在接近平直的曲线；例如，西班牙裔的收入低于白人，几乎在所有的分位数上都是相似的，这使得曲线是平直的。然而，大多数协变量不仅产生位置变化，而且导致显著的形状变化。

图 7.4 是关于对数系数的图像。我们发现：对数转换的近似正态性缩短了右偏分布。因此，常数系数的图像类似于正态分布的分位数函数。正如第 4 章讨论的，对数系数以相对形式模拟成比例的变化；水平直线表示在不改变偏态的情况下的位置变化和尺度变化。解释任何背离水平直线的情况都是比较困难的，因为它可能暗示着位置、尺度和偏态变化的混合情形。另外，因为在对数单位情况下，高端分位数水平直线之上或之下的对数收入的微量变化会造成初始收入的巨大变化，所以在声称曲线为"接近平直"(close-to-flat)时，我们应该谨慎。例如，三个最低教育水平组的曲线相当平直，但我们不会称它们为"接近平直"，因为它们在第 0.8 分位数之上的上尾部分是明显下降的。简言之，与初始单位系数的图像相比，对数系数的图像是较无效的，而且在解释时要更加谨慎。

非中心位置的分位数回归：绝对效应

图像视角提供了关于协变量影响条件收入分布形状的总的看法。我们现在仔细观察非中心位置的情况，以补充图像视角的不足。我们选择超出图像之外的两个刚刚检验过的极端值：第 0.025 和第 0.975 分位数。为了获得初始单位收入在第 0.025 和第 0.975 分位数回归方程的系数标准误，我们可以用 500 个复制样本的"sqreg"命令，或者手工执行 500 个复制样本的自举法，以保存计算系数估计值的所有 500 组情况。条件形状变化分位数的计算是以涉及这两个分位数估计值（第 0.025 和第 0.975）的每一个自举样本为基础的，所以我们在

这里展示的是手工得到的自举结果。在 500 组系数估计值中，我们使用中位数作为点估计值和 95%置信区间。如果这一置信区间没有穿过 0 值，那么这一系数在 $p = 0.05$ 处是显著的。这些结果几乎等同于 sqreg 的结果。

对数收入方程的估计值不是以绝对形式出现的。因为绝对效应对于理解协变量对分布形状的影响作用至关重要，所以我们需要找到绝对效应，这可在典型设定（所有协变量的均值）下计算得到。至于获得初始收入，我们保存了从自举样本中得到的 500 组对数单位系数。对于以自举样本为基础的估计方法中的每一个协变量，我们进行如下处理：

(1) 通过与常数项相加，获得协变量均值的单位增量下的对数条件分位数。

(2) 分别对对数条件分位数和常数项取指数，获得两个初始单位下的条件分位数。

(3) 求这两个初始单位条件分位数的差，它代表了从典型设定（TSE）中求得的协变量的绝对效应。

表 7.5 展示了收入和对数收入在第 0.025 和第 0.975 分位数处的绝对效应。表 7.5 的上半部分数值来自收入方程。常数项分别表示当所有协变量取均值时第 0.025 和第 0.975 分位的估计值：最低值大约为 1 万美元，最高值大约为 13.7 万美元。

最突出的模式是协变量数值两端的效应存在巨大的差距。例如，黑人的收入在第 0.025 分位数处下降了 1991 美元，而在第 0.975 分位数处则减少了 17380 美元。另外，西班牙人和亚洲人在第 0.025 分位数处相比于白人有显著的较低收入，而在第 0.975 分位数下则不存在差别。

表 7.5 的下半部分展示了基于对数收入方程的典型设定效应（TSE）。常数项表示在典型设定下的第 0.025 和第 0.975 条件分位数。TSE 系数近似于从收入方程估计得到的系数。但它们并不完全相等，因为对数收入模型的拟合优于收入模型，还因为对数收入方程的估计值是在典型设定下求得的。

表 7.5 尾端分位数的绝对效应(2001 年的初始和对数收入)

变量	第 0.025 分位数	第 0.975 分位数
	系数	系数
收入模型		
年龄	248**	3103**
年龄平方	−2**	−29**
黑人	−1991**	−17380**
西班牙人	−2495**	−7418
亚洲人	−4221**	16235
大学未毕业	−2607**	−105858**
高中毕业	−4332**	−119924**
高中以下	−6211**	−129464**
已婚无子女	−4761**	−18878**
单亲母亲	−10193**	−50465**
单身	−12257**	−78570**
其他类型	−7734**	−16876**
农村	−943**	−18654**
常数	10156**	137561**
对数收入模型		
年龄	396**	5409**
年龄平方	−3**	−53**
黑人	−2341**	−28867**
西班牙人	−1835**	−8032
亚洲人	−3259**	8636
大学未毕业	−1916**	−49898**
高中毕业	−2932**	−57557**
高中以下	−4095**	−70006**
已婚无子女	−3149**	−12471**
单亲母亲	−5875**	−33219**
单身	−6409**	−63176**
其他类型	−4382**	−5282**
农村	−938**	−26742**
常数	8457**	115804**

评估影响位置和形状变化的协变量效应

QRM 估计值可以被用作精确计算协变量是如何改变条件分布的位置和形状的。为了完成这项估计，我们比较两个组:参照组和比较组。在连续型协变量的情况下，当保持其他协变量固定不变时，通过赋予协变量某些数值来定义参照组，而通过增加协变量的一个单位来定义比较组。对于二元协变量而言，在其他协变量保持不变时，我们将其数值从 0 换成 1。为了反映初始单位的分布，所有的比较都在绝对形式下进行。因此，如果使用对数收入回归方程来拟合数据，首先要获得绝对形式下的协变量系数(如之前部分所提到的)。位置的变化可通过中位数的系数得到。而形状(尺度和偏态)的变化则是在多个系数组合的基础上获得的。它们的显著水平可通过自举方法确定。

表 7.6 展示了 1991 年和 2001 年收入模型的结果，位置变化结果在顶部，尺度变化结果在中间，而偏态变化结果在底部。在 1991 年，除亚洲人之外的所有协变量都显著地改变了比较组相对于参照组的位置。其中一些效应从 1991 年到 2001 年发生了明显的变化。亚洲人的位置变化在 1991 年是不显著的，而在 2001 年则显著为负，意味着白人在收入上相对于少数族裔享受着绝对优势。然而，其他种族/民族组的位置变化是不明显的。年龄的位置变化在 2001 年比 1991 年变得更不重要了。这种情况同样存在于低教育水平上。然而，已婚有子女家庭除外的家庭类型的负向位置变化效应更强，正如农村变量的效应一样。

位置变化代表组间差别。正如之前对表 7.4 讨论的，中位数回归系数比 OLS 系数更弱。对于高度右偏的收入分布，中位数回归系数代表中心位置的变化，而 OLS 系数则更多地受到右尾的影响。通过位置变化(中位数回归)，我们关于教育组别的发现意味着教育在位置变化方面的效应并不如文献中指出的那样强。对位置变化的影响，或者组间差别，只是 10 年间不平等如何变化这个故事中的一部分;另一部分则是形状变化，或者相对的组内差别。QRM 的优势在于，它们区分了组间和组内的差别，增加了我们对不平等变化的理解。

表 7.6 条件分位数的位置和形状变化:来自初始单位的 QRM

变　化	1991	2001
位置变化		
年龄	1801**	1501**
年龄平方	−169**	−149**
黑人	−7878**	−7473**
西班牙人	−8692**	−7616**
亚洲人	−1231	−2850**
大学未毕业	−19173**	−18588**
高中毕业	−25452**	−24926**
高中以下	−32595**	−30345**
已婚无子女	−9562**	−11501**
单亲母亲	−22366**	−25862**
单身	−27866**	−32039**
其他类型	−11716**	−13659**
农村	−5284**	−6698**
尺度变化		
(总体中间的 95%)		
年龄	3393**	2852**
年龄平方	−305**	−272**
黑人	−14617**	−15378**
西班牙人	−3027	−4893
亚洲人	11425	20842
大学未毕业	−34212**	−103245**
高中毕业	−49002**	−115600**
高中以下	−63477**	−123369**
已婚无子女	3708	−14001**
单亲母亲	−9177	−40290**
单身	−32482**	−66374**
其他类型	−8220	−8819**
农村	−9817**	−17693**
偏态变化		
(总体中间的 95%)		
年龄	−0.0200**	−0.0195**
年龄平方	0.0003**	0.0002**
黑人	0.0242	0.0713
西班牙人	0.2374**	0.1833**

（续表）

变　化	1991	2001
亚洲人	0.0395	0.1571
大学未毕业	0.3524**	−0.8572
高中毕业	0.5245**	−1.0263
高中以下	0.7447**	−1.1890
已婚无子女	0.4344**	0.1514
单亲母亲	0.8493**	0.3781**
单身	0.5229**	0.2184
其他类型	0.1748	0.1714
农村	0.0446	0.0541

尺度变化是形状变化的一种类型。在3个少数族裔组中，只有黑人的条件收入分布范围比白人的短。黑人中95%的收入范围比白人的更窄，意味着黑人群体比白人群体有着更高的同质性，而且种族在决定收入时具有显著性。这一尺度变化在2001年变得更大。同样的情况存在于3个低教育组别中。教育的尺度变化提供了关于在收入决定因素中教育重要性不断增加的一致的和精确的发现：形状变化而不是位置变化，表明教育的重要性在不断增加。

偏态变化是形状变化的另一种类型。条件分位数偏态程度的增加预示着不平衡的组内差别，有利于分布顶部的成员。1991年的结果表明：许多处于不利地位的组别都经历了这种不平衡的组内差别，包括西班牙人、3个低教育水平组和处于劣势的家庭类型（单亲母亲、单身和其他类型家庭）。其中一些组内差异在2001年消失了，特别是在教育方面。这一发现进一步反映了社会奖励大学毕业生和限制低教育群体中非常能干者向上流动的机制。

表7.7展示了从对数收入模型中求得的初始单位结果。这些结果反映了生命周期中的相同趋势，种族/民族组别、教育组别、家庭类型和农村居住类型。不管拟合的是收入还是对数收入，在每一年和年代的趋势中，位置变化和尺度变化是相似的。偏态变化则存在一些差异。特别的，2001年，低教育组的偏态程度显著减小了；但这一发现在对数收入模型上是显著的，而在收入模型上则是不显著的。当检验两种模型（收入和对数收入）的拟合情况时，出现这种矛盾

并不奇怪。它们体现了两种根本上不同的模型,其中一种(对数收入)模型的拟合情况更优。另一方面,如果定性的结果是不一样的,这可能说明这些结果是不稳健的。我们需要通过观察协变量影响不平等程度的作用的综合评价,来决定以上情况是否属实。

表 7.7 条件分位数的位置和形状变化(来自对数单位的 QRM)

变　化	1991	2001
位置变化		
年龄	2456**	1994**
年龄平方	−24**	−20**
黑人	−9759**	−8386**
西班牙人	−7645**	−6300**
亚洲人	−1419	−3146**
大学未毕业	−10635**	−11012**
高中毕业	−14476**	−15485**
高中以下	−20891**	−20892**
已婚无子女	−3879**	−5103**
单亲母亲	15815**	−17506**
单身	−19599**	−21658**
其他类型	−6509**	−7734**
农村	−4931**	−6725**
尺度变化		
(总体中间的 95%)		
年龄	4595**	5008**
年龄平方	−41**	−50**
黑人	−17244**	−26509**
西班牙人	−2503	−6017
亚洲人	4290	12705
大学未毕业	−22809**	−47992**
高中毕业	−32675**	−54434**
高中以下	−44457**	−65956**
已婚无子女	77	−9264**
单亲母亲	−10269	−27272**
单身	−32576**	−56791**
其他类型	−7535	−906
农村	−12218**	−25760**

（续表）

变　化	1991	2001
偏态变化		
(总体中间的 95%)		
年龄	−0.0417**	−0.0100
年龄平方	0.0005**	0.0002
黑人	0.1127	−0.0682
西班牙人	0.2745**	0.1565**
亚洲人	−0.0383	0.1469
大学未毕业	0.0655	−0.2775**
高中毕业	0.0934	−0.2027**
高中以下	0.2742**	−0.1456**
已婚无子女	0.0890	−0.0272
单亲母亲	0.5404**	0.3193**
单身	0.2805**	−0.0331
其他类型	0.0164	0.1640**
农村	0.0012	−0.0740

我们发展出对协变量影响不平等程度的作用的综合评价方法，它用于检验位置和形状变化符号的组合情况。

我们只考虑显著的变化情况。对于协变量而言，在这三种变化中同步的符号表明协变量加剧了不平等；显著的符号的个数越多，加剧效应就越强。非同步符号则表明协变量可能在增加组别不平等的同时降低了组内不平等，反之亦然。表 7.8 中关于收入模型的左边部分表明：在 1991 年，任何一个协变量对不平等都不具有同步效应，但在 2001 年，许多协变量都具有这一效应。这些同步协变量包括教育组别、家庭类型（单亲母亲除外）和农村居住类型。右边两列展示了对数收入模型的相应结果。我们发现：在综合评价上不存在明显的差别。例如，教育组别在两种模型中的效应模式在 1991 年是非同步的，而到 2001 年则转变为同步的。因此，2001 年的美国社会是更加不平等的，而且它的社会分层比 10 年前更多地受教育、婚姻、有否子女和农村居住的影响。

在这个例子中，我们使用总体中 95%的个案计算形状变化的分位数。研究者可以根据他们的研究问题定义自己关心的形状变化。它可以将相应的形状

变化定为总体中间的 99%、98%、90%、80%或者 50%部分。我们将这个任务留给读者自己完成。

表 7.8 协变量对不平等程度影响效应的综合评价(系数的同步性模式)

变量	收入方程		对数收入方程	
	1991	2001	1991	2001
年龄	+ + −	+ + −	+ − −	+ + 0
年龄平方	− − +	− − +	− − +	− − 0
黑人	− − 0	− − 0	− − 0	− − 0
西班牙人	− 0 +	− 0 +	− 0 +	− 0 +
亚洲人	0 0 0	− 0 0	0 0 0	− 0 0
大学未毕业	− − +	− − 0	− − 0	− − −
高中毕业	− − +	− − 0	− − 0	− − −
高中以下	− − +	− − 0	− − +	− − −
已婚无子女	− 0 +	− − 0	− 0 0	− − 0
单亲母亲	− 0 +	− − +	− 0 +	− − +
单身	− − +	− − 0	− − +	− − 0
其他类型	− 0 0	− − 0	− 0 0	− 0 +
农村	− − 0	− − 0	− − 0	− − 0

小结

最近 10 年收入不平等维系和扩大的根源是什么？为处理这一研究问题，我们使用了本书发展出的技术。首先，我们用第 2 章介绍的分位数概念进行描述性分析。对收入数据，我们讨论了右偏分布和顶端编码的问题，并解释 QRM 为什么可以和怎样共同解决这些问题。我们的分析按照第 3 章讨论的 6 个步骤进行：定义和拟合模型，评估拟合优度，计算参数的推论统计，图像化系数和它们的置信区间，计算位置和形状变化以及它们的推论统计值。我们分别描述了收入和对数收入模型，特别关注重建初始单位系数。在描述这些步骤的同时，我们通过对结果的解释，展示了用 QRM 技术处理研究问题时所具有的实用性。我们希望关于应用步骤的系统总结为实证研究提供清晰的指导。

附录 | STATA 命令

数据：d0.dta 是一个用于分析的 Stata 文件

Ⅰ. 分析初始单位收入的 Stata 命令

第 1 步:拟合优度

```
* q0.do
* a full model
* raw-scale income in $1000
* OLS
* 19 quantiles

tempfile t

use d0

global X age age2 blk hsp asn scl hsg nhs mh fh sg ot rural
* centering covariates
sum $X
tokenize $X
  while "'1 '"~ = ""{
    egen m = mean ('1 ')
    replace '1m = '1 '—m
    drop m
```

```
    macro shift
}
sum $X

forvalues k = 1/2 {
reg cinc  $X if year = 'k '
}

forvalues i = 1/19 {
local j = 'iS/20
qreg cinc  $X if year = 1, q('j ') nolog
}

forvalues i = 1/19 {
local j = 'i '/20
qreg cinc  $X if year = 2, q('j ') nolog
}
```

第 2 步:500 次重复抽样的同时分位数回归(Simultaneous Quantile Regressions)

```
s0.do
full model
* sreq 19 quaniles
* raw-scale income in $1000
analysis for 2001

tempfile t
set matsize 400
global X age age2 blk hsp asn scl hsg nhs mh fh sg ot rural
```

```
use cinc $X year if year = 2 using do, clear
drop year

*centering covariates
sum $X
tokenize $X
  while '''1 '''~ = ''''{
    egen m = mean ('1 ')
    replace '1 ' = '1 '—m
    drop m
    macro shift
}
sum $X

sqreg cinc $X, reps(500) q(.05 .10 .15 .20 .25 .30 .35 .40 .45 .50 .55 .60
.65 .70 .75 .80 .85 .90 .95)
mstore b, from(e(b))
mstore v, from(e(V))
keep age
keep if—n<11
save S0, replace
```

第 3 步:基于 s0. do 的结果创建图表

```
* s_m0 .do
* matrix operation
* 13 covariates + cons
* graphs for beta 's (19 QR)
* 500 bootstrap se
```

```
* analysis for 2001

* for black-white graphs
set scheme s2mono

set matsize 400

* 13 covariate + cons
local k = 14
* k parameters for each of the 19 quantiles
local k1 = 'k ' * 19

use s0, clear

qui mstore b
qui mstore v

* 95 % ci
* dimension 'k 'x 1

mat vv = vecdiag(v)
mat vv =  vv '
svmat vv
mat drop vv
qui replace w1 = sqrt(vv1)
mkmat vv1 if _n< = 'k1 ', mat(v)
drop vv1
```

```
mat b = b '
mat 1 = b - 1.96 * v
mat u = b + 1.96 * v

 * 19 quantiles
mat
q = (0.05\0.10\0.15\0.20\0.25\0.30\0.35\0.40\0.45\0.50\0.55\0.60\0.65
\0.70\0.75\0.80\0.85\0.90\0.95)

 * reorganize matrix by variable

forvalues j = 1/'k '{
forvalues i = 1/19 {
local 1 = 'k ' * ('i ' - 1) + 'j '
mat x 'j 'q 'i ' = q['i ', 1], b['1 ', 1], 1['1 ', 1], u['1 ', 1], v['1 ', 1]
}
}

forvalues j = 1/'k '{
mat x 'j ' = x 'j 'q1
forvalues i = 2/19 {
mat x 'j ' = x 'j '\x 'j 'q 'i '
}
* q b l u v
mat list x 'j ', format( %8.3f)
svmat x 'j '

mat al = x 'j '[1 ..., 2]
```

```
mat a2 = x 'j1[1 ..., 5]
mat xx 'j ' = q, al, a2
 * q b v

mat list xx 'j ', format( %8.3f)
mat drop al a2 xx 'j '
}

 * graphs using the same scale for categorical covariates
 * use age, age-squared and constant as examples
 * age
twoway rarea x13 x14 x11, color(gsl4) || line x12 x11, lpattern(solid)
yline(0,
lpattern(solid) lwidth(medthick)) ylabel(0 "0" 1 "1000" 2 "2000" 3
"3000")
ytitle(quantile coefficients for income ( $) ) xtitle(p) xlabel(0(.1)1)
legend (off )
graph export g0.ps, as(ps) logo(off) replace

 * age2
twoway rarea x23 x24 x21, color(gsl4) || line x22 x21, lpattern(solid)
yline(0.
lstyle(foreground) lpattern(solid) lwidth(medthick)) xtitle(p) xlabel(0
(.1)1)
legend(off)
graph export g2.ps, as(ps) logo(off) replace

 * constant (the typical setting)
```

```
twoway rarea x143 x144 x141, color(gsl4) || line x142 x141, lpattern
(solid)
yline(0, lstyle(foreground) lpattern(solid) lwidth(medthick)) ylabel(0
(20)120)
xlabel(0(.1)1) xtitle(p) legend(off)
graph export gl4.ps, as(ps) logo(off) replace

drop x*
matrix drop _all
```

第 4 步：计算位置和形状变化

```
* e0.do
* full model
* raw-scale income in $1000
* bootstrap
* analysis for 2001

tempfile t

global X age age2 blk hsp asn scl hsg nhs mh fh sg ot rural

use cinc SX year if year = 2 using do, clear
drop year

* centering covariates
sum $X
tokenize $X
  while '''1 '''~ = ''''{
```

```
    egen m = mean ('l ')
replace 'l ' = 'l '—m
drop m
macro shift
}

sum $X
save 't '
forvalues i = 1/500 {
use 't ', clear
bsample
qreg cinc SX, q(.025) nolog
mstore e, from(e(b))
keep if _n<11
keep age
save e0 'i ', replace
}
```

[修改 e0.do 中的命令从而生成分析第 0.5 分位数的 el.do 和分析第 0.975 分位数的 e2.do]

```
* bs0.do
* location and shape shift quantities
* bootstrap confidence interval
* 3 quantiles (.025, .5, .975)

set matsize 800
```

```
* k= #of covariates + cons
local k=14
local kl='k'—1

* initial
forvalues j=0/2 {
use e'j'l, clear
qui mstore e
mat ren e e'j'
}

forvalues j=0/2 {
forvalues i=2/500 {
use e'j''i', clear
qui mstore e
mat e'j'=e'j'\e
mat drop e
}

forvalues j=0/2 {
qui svmat e'j'
}

* mean of estimate (point estimate)
* percentile-method (95% ci)
forvhlues j=0/2 {
forvalues i=1/'k'{
pctile x=e'j''i', nq(40)
```

```
sort x
qui gen x0 = x if _n = 20
qui gen xl = x if _n = 1
qui gen x2 = x if _n = 39
egen em 'j ''i ' = max(x0)
egen el 'j ''i ' = max(xl)
egen eu 'j ''i ' = max(x2)
drop x xo xl x2
sum em 'j ''i 'el 'j ''i 'eu 'j ''i '
}
}

* SCS scale shift
forvalues i = 1/'kl '{
gen scls 'i ' = e2 'i ' - e0 'i '
pctile x = scls 'i ', nq(40)
sort x
qui gen x0 = x if _n = 20
qui gen xl = x if _n = 1
qui gen x2 = x if —n = 39
egen sclsm 'i ' = max(x0)
egen sclsl 'i ' = max(xl)
egen sclsu 'i ' = max(x2)
drop x x0 xl x2
sum sclsm 'i 'sclsl 'i 'sclsu 'i '
}

* SKS skewedness shift
```

```
* SKS e2(.975) - e1(.5) and e1(.5) - e0(.025)
* i for covariate, k for constant
forvalues i=1/'kl'{
gen nu=(e2'i'+e2'k'-e1.i'-el'k')/(e2'k'-el'k')
gcn de=(el'i'+el'k'-e0'i'-e0'k')/(el'k'-e0'k')
gcn skls'i'=nu/de
drop nu de
pctile x=skls'i', nq(40)
sort x
qui gen x0=x if _n=20
qui gen xl=x if _n=I
qui gen x2.x if _n=39
egen sklsm'i'=max(x0)
egen sklsl'i'=max(xl)
egen skleu'i'=max(x2)
drop x xo xl x2
sum sklsm'i'sklsl'i'sklsu'i'
}
```

Ⅱ. 分析对数单位收入的 Stata 命令

[将初始单位收入替换为对数单位收入,重复第 1、2、3 步]

第 4 步: 基于对数收入 QRM 计算初始位置和形状变化

[将 e0.do、el .do 和 e2.do 文件中的初始单位收入替换为对数单位收入]

```
set matsize 800
* k= # of covariates + cons
local k=14
```

```
local k1 = 'k ' - 1

 * parameter matrix (e0 e1 e2)
 * initial
forvalues j = 0/2 {
use e 'j '1, clear
qui mstore e
mat ren e e 'j '
}

 * 500 reps
forvalues j = 0/2 {
forvalues i = 2/500 {
use e 'j ''i ', clear
qui mstore e
mat e 'j ' = e 'j '\e
mat drop e
}
}

get log conditional quantile
forvalues j = 0/2 {

 * dimensions 500 x 14
 * c 'j '1 to c 'j '13 are covariates
 * c 'j '14 constant

forvalues m = 1/'k '{
```

```
mat c 'j ''m ' = e 'j '[1 . . . , 'm '1
I
forvalues m = 1/'kl '{
mat c 'j ''m ' = c 'j ''m ' + c 'j ''k '
}

mat c 'j ' = c 'j '1
mat drop c 'j '1
forvalues m = 2/'k '{
mat c 'j ' = c 'j ', c 'j ''m '
mat drop c 'j ''m '
}

* transform log-scale conditional quantile to raw-scale conditional quan-
tile
* matrix to var
svmat c 'j '
mat drop c 'j '
forvalues m = 1/'k '{
qui replace c 'j ''m ' = exp(c 'j ''m ')
}

forvalues m = 1/'kl '{
m i replace c 'j ''m ' = c 'j ''m ' - c 'j ''k '
}

* var to matrix
forvalues m = 1/'k '{
```

```
mkmat c'j''m', mat(e'j''m')

}

mat e'j'=e'j'1
mat drop e'j'1
forvalues m=2/'k'{
mat e'j'=e'j', e'j''m'
mat drop e'j''m'
}
mstore e'j', from(e'j') replace
}

mat dir
keep age
keep if _n<11
save 1-r, replace

***

* bsl.do
* bootstrap method
* location and shape shift quantities
* based on log-to-raw coeff

set matsize 800

* k= # of covariates + cons
```

```
local k = 14
local kl = 'k ' - 1

use 1 - r

forvalues j = 0/2 {
qui mstore e 'j '
qui svmat e 'j '
}

 * mean of estimate (point estimate)
 * sd of estimates (se)
 * percentile-method (95 % ci)
forvalues j = 0/2 {
forvalues i = 1/'k '{
pctile x = e 'j ''i ', nq(40)
sort x
qui gen x0 = x if  - n  = 20
qui gen xl = x if  - n = 1
qui gen x2 = x if  - n = 39
egen em 'j ''i ' = max(x0)
egen el 'j ''i ' = max(xl)
egen eu 'j ''i ' = max(x2)
drop x xo xl x2
sum em 'j ''i 'el 'j ''i 'eu 'j ''i '
}
}
```

```
* scs scale shift
forvalues i = 1/'kl '{
gen scls 'i ' = e2 'i ' - e0 'i '
pctile x = scls 'i ', nq(40)
sort x
qui gen x0 = x if  - n  = 20
qui gen xl = x if  - n = 1
qui gen x2 = x if  - n = 39
egen sclsm 'i ' = max(x0)
egen sclsl 'i ' = max(xl)
egen sclsu 'i ' = max(x2)
drop x xo xl x2
sum sclsm 'i 'sclsl 'i 'sclsu 'i '
}

* SKS skewedness shift
* SKS e2(.975)  -  el(.5) and el(.S)  -  e0(.025)
* i for covsriate, k for constant
forvalues i = 1/'kl '{
pen nu = (e2 'i ' + e2 'k ' - el 'i ' - el 'k ')/(e2 'k ' - el 'k ')
gen de = (el 'i ' + el 'k ' - e0 'i ' - e0 'k ')/(el 'k ' - e0 'k0)
gen skls 'i ' = nu/de
drop nu de
pctile x = skls 'i ', nq(40)
sort x
qui gen x0 = x if  - n = 20
qui gen xl = x if  - n = 1
qui gen x2 = x if  - n = 39
```

```
egen sklsm'i'=max(x0)
egen sklsl'i'=max(xl)
egen sklsu'i'=max(x2)
drop x xo xl x2
sum sklsm'i'sklsl'i'sklsu'i'
}
```

注释

[1] 为了准确起见，我们假设样本大小是奇数。如果样本大小是偶数的，那么样本中位数则被定义为第 $(n/2)$ 位和第 $\left(\frac{n}{2}+1\right)$ 位顺序统计值的平均值，在修改位于第 $(n/2)$ 位 $\left[\text{或第} \left(\frac{n}{2}+1\right) \text{位}\right]$ 顺序统计值之上（或之下）的一个数据值同时保持它的相对位置不变时，这一表述依然正确。

[2] 这里用到的数据来自收入与项目参与调查(SIPP)的 2001 年面板数据。家庭收入是指 2001 年的年收入。第 3 章至第 5 章中用到的分析样本包括 19390 户白人家庭和 3243 户黑人家庭。

[3] $Q^{(q)}(y_i \mid x_i) = Q^{(q)}(\beta_0^{(p)} + x_i\beta_1^{(p)} + \varepsilon_i^{(p)}) = \beta_0^{(p)} + x_i\beta_1^{(p)} + Q^{(q)}(\varepsilon_i^{(p)}) = Q^{(q)}(y_i \mid x_i) + c_{p+q}$。

[4] 然而，不同分位数解决方法的数量受限于有限样本的大小。

[5] 准确地说，百分比的变化是 $100(e^{0.115} - 1) = 12.2\%$。

[6] 条件均值是与线性预测变量的指数形式成比例的(Manning, 1998)。例如，如果误差服从正态分布 $N(0, \sigma_\varepsilon^2)$，那么 $E(y_i \mid x_i) = e^{\beta_0 + \beta_1 x_i + 0.5\sigma_\varepsilon^2}$。$e^{0.5\sigma_\varepsilon^2}$ 有时被称为拖尾因子(smearing factor)。

[7] 注意稳健性不适用于协变量的离群值。

[8] 在 QRM 中，我们假设 $\varepsilon^{(p)}$ 的第 p 分位数等于 0。

[9] 参照/比较这对术语可见 Handcock and Morris (1999)。

[10] 我们可以谈论增加一年教育的效应，而这对于所有种族和所有教育水平而言是一样的。相似的，从黑人变为白人也存在一种影响效应，对于所有教育水平来说也是相同的。这里存在黑人变为白人的效应，与白人变为黑人的效应相反。对没有交互项的 LRM 的位置效应的分析是十分简单的。当模型引入交互项时，这一分析便变得相当复杂。

[11] 需要注意的是，我们可以指定任何分位数，例如第 0.39 分位数，而不局限在等距分位数上。

[12] 如果估计系数是 $\hat{\beta}$，那么预测变量的单位增量会使因变量增加 $[100(e^{\hat{\beta}} - 1)]\%$，对于估计系数 $\hat{\beta}$ 的小数值而言，这大概等于 $100\hat{\beta}\%$。

[13] 这些实践包括从 logit, probit 和 tobit 模型估计对概率的影响效应。

参考文献

Abreveya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth oucomes. *Empirical Economics*, *26*, 247—257.

Austin, P., Th, J., Daly, P., & Alter, D. (2005). The use of quantile regression in health care research: A case study examining gender differences in the timeliness of thrombolytic therapy. *Statistics in Medicine*, *24*, 791—816.

Bedi, A., & Edwards, J. (2002). The impact of school quality on earnings and educational returns—evidence from a low-income country. *Journal of Development Economics*, *68*, 157—185.

Berry, W. D. (1993). *Understanding regression assumptions*. Newbury Park, CA: Sage Publications.

Berry, W D., & Feldman, S. (1985). *Multiple regression in practice*. Beverly Hills, CA: Sage Publications.

Buchinsky, M. (1994). Changes in the U. S. wage structure 1963—1987: Application of quantile regression. *Econometrica*, *62*, 405—458.

Budd, J. W., & McCall, B. P. (2001). The grocery stores wage distribution: A semi-parametric analysis of the role of retailing and labor market institutions. *Industrial and Labor Relations Review*, *54*, *Extra Issue*: *Industry Studies of Wage Inequality*, 484—501.

Cade, B. S., Terrell, J. W., & Schroeder, R. L. (1999). Estimating effects of limiting factors with regression quantiles. *Ecology*, *80*, 311—323.

Chamberlain, G. (1994). Quantile regression, censoring and the structure of wages. In C. Skins (Ed.), *Advances in Econometrics* (pp. 171—209). Cambridge, UK: Cambridge University Press.

Chay, K. Y., & Honore, B. E. (1998). Estimation of semiparametric censored regression models: An application to changes in black-white earnings inequality during the 1960s. *The Journal of Human Resources*, *33*, 4—38.

Edgeworth, F. (1888). On a new method of reducing observations relating to several quantiles. *Philosophical Magazine*, *25*, 184—191.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1—26.

Eide, E. R., & Showalter, M. H. (1999). Factors affecting the transmission of earnings across generations: A quantile regression approach. *The Journal of Human Resources*, *34*, 253—267.

Eide, E. R., Showalter, M., & Sims, D. (2002). The effects of secondary school quality on the distribution of earnings. *Contemporary Economic Policy*, *20*, 160—170.

Feiring, B. R. (1986). *Linear programming*. Beverly Hills, CA: Sage Publications. Fortin, N.

M., & Lemieux, T. (1998). Rank regressions, wage distributions, and the gender gap. *The Journal of Human Resources*, *33*, 610—643.

Handcock, M. S., & Morris, M. (1999). *Relative distribution methods in the social sciences.* New York: Springer.

Hao, L. (2005, April). *Immigration and wealth inequality: A distributional approach.* Invited seminar at The Center for the Study of Wealth and Inequality, Columbia University.

Hao, L. (2006a, January). *Sources of wealth inequality: Analyzing conditional distribution.* Invited seminar at The Center for Advanced Social Science Research, New York University.

Hao, L. (2006b, May). *Sources of wealth inequality: Analyzing conditional location and shape shifts.* Paper presented at the Research Committee on Social Stratification and Mobility (RC28) of the International Sociological Association (ISA) Spring meeting 2006 in Nijmegen, the Netherlands.

Kocherginsky, M., He, X., & Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, *14*, 41—55.

Koenker, R. (1994). Confidence intervals for regression quantiles. In *Proceedings of the* 5th *Prague symposium on asymptotic statistics* (pp. 349—359). New York: Springer-Verlag.

Koenker, R. (2005). *Quantile regression.* Cambridge, UK: Cambridge University Press.

Koenker, R., & Bassett, Jr., G. (1978). Regression quantiles. *Econometrica*, *46*, 33—50.

Koenker, R., & d'Orey, V. (1987). Computing regression quantiles. *Applied Statistics*, *36*, 383—393.

Koenker, R., & Hallock, K. F. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives*, *15*, 143—156.

Koenker, R., & Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of Econometrics*, 93, 327—344.

Lemieux, T. (2006). Post-secondary education and increasing wage inequality. *Working Paper No. 12077.* Cambridge, MA: National Bureau of Economic Research.

Machado, J., & Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, *20*, 445—465.

Manning, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, *17*, 283—295.

Melly, B. (2005). Decomposition of differences in distribution using quantile regression. *Labour Economics*, *12*, 577—590.

Mooney, C. Z. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* Newbury Park, CA: Sage Publications

Scharf, F. S., Juanes, F., & Sutherland, M. (1989). Inferring ecological relationships from the edges of scatter diagrams: Comparison of regression techniques. *Ecology*, *79*, 448—460.

Scheffé, H. (1959). *Analysis of variance.* New York: Wiley.

Schroeder, L. D. (1986). *Understanding regression analysis: An introductory guide*. Beverly Hills, CA: Sage Publications.

Shapiro, I., & Friedman, J. (2001). *Income tax rates and high-income taxpayers: How strong is the case for major rate reduction?* Washington, DC: Center for Budget and Policy Priorities.

U. S. Census Bureau. (2001). *U. S. Department of Commerce News*. (CB01-158). Washington, DC.

Wei, Y., Pere, A., Koenker, R., & He, X. (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, *25*, 1369—1382.

译名对照表

conditional mean	条件平均数
conditional-mean function	条件均值函数
heteroscedastic error	异方差误差模型
conditional-mean models	条件均值模型
conditional scale	条件单位
conditional-mean modeling framework	条件均值的建模框架
noncentral locations	非中心位置
median regression	中位数回归
least-absolute-distance estimation	最小绝对距离估计
least squares estimation	最小二乘估计
quantile-based measures	分位差方法
minimum distance principle	最小距离原则
monotonic equivariance principle	单调同变性
asymptotic inference	渐近推断
bootstrap method	拔靴法
distributional interpretation	分布理解
scale shift	尺度变化
inverse	逆反函数
empirical or sample cdf	经验或样本分布函数
sampled distribution	抽样分布
sample quantiles	样本分位数
derivative	导数
scale and skewness	刻度和偏态
quantile-based scale measure	分位差刻度测量
moment-based measure	动差法
upper spread	上端部分
quantile-based skewness(QSK)	分位差偏态测量法
average squared deviation	均方差
mean absolute distance	平均绝对距离
sample points	样本点
mean squared deviation	均值平方差
monotone equivariance	单调同变性
quantile-based procedures	分位差程序
quantile-regression model(QRM)	分位数回归模型
one-model assumption	单一模型假设

conditional median regression	条件中位数回归
the sum of absolute residuals	绝对残差总和
median-regression line	中位数回归线
point/line duality	点-线二元性
exterior-point algorithms	外点计算
polyhedral surface	多面体曲面
linear equivariance	线性同变性
smearing factor	拖尾因子
asymptotic procedure	渐近程序
scalar multiple	纯量倍数
multivariate normal approximation	多变量正态近似法
corresponding diagonal element	相应对角元素
bootstrap sample	拔靴样本
reference and comparison	参考与比较
quantile-based skewness	分位差偏态
skewness shift	偏态变化
in-sync	同步模式
out of sync	不同步模式
Studentized Range Test	差距检定法
typical-setting effects, TSE	典型设定值
mean effect, ME	均值效应
confidence envelope	置信封闭间
top-coded	顶端编码

空间回归模型

作 者 简 介

迈克尔·沃德

迈克尔·沃德(Michael D. Ward)是华盛顿大学西雅图分校政治科学系教授(http://faculty.washington.edu/mdw),讲授政治学研究方法和国际关系课程,同时是统计和社会科学中心成员。目前的工作包括研究国际贸易和冲突的动态网络中的依赖关系。此外,他还关注波斯尼亚以及俄罗斯北高加索地区不同国家团体之间的社会距离。

克里斯蒂安·格里蒂奇(Kristian Skrede Gleditsch)是埃塞克斯大学政府系教授(http://privatewww.essex.ac.uk/～ksg)以及奥斯陆国际和平研究院(PRIO)内战研究中心的研究员。其研究兴趣包括冲突与合作,民主化和社会与政治过程中的空间维度。

译 者 简 介

宋曦

宋曦2008年毕业于中国人民大学社会学系,2010年毕业于香港科技大学社会科学部,获硕士学位,目前是加州大学洛杉矶分校(UCLA)社会学系博士研究生。

序

社会科学定量研究的很多方法都是基于对个体行为的分析。我们通常将个体行为作为结果变量，然后将其表示成由一组解释变量构成的方程，最常见的就是回归方程的形式。社会科学理论主要就是描述这种结果变量和解释变量之间是如何关联的。当然，这种分析不仅停留在个体或者微观层面。有时，研究者也会分析集体层面的数据，比如街道、社区、公司、城市、县、州以及国家。但是分析的逻辑仍然相同。我们试图构造因变量和某些自变量（诸如个人、街坊、社区、公司、城市、县、州和国家）之间的关联，不管这种关联是否为因果关系。当我们这样做的时候，其实已经暗示了观测值的地理或者空间位置并不起作用。尽管研究者也会经常用虚拟变量将处于同一个地区的观测值合成一组，但这样做并没有考虑除了空间以外的其他相似性。比如，数据分析者常用虚拟变量来划分个体是来自美国南部还是其他地区。这样做是为了控制一些特有的文化特征，而不是出于对回归中空间依赖关系的考虑。空间回归模型一书将不遗余力地解决线性回归分析中空间依赖关系的相关问题。

确切地说，我们将结果变量 y_i 对解释变量的向量 x_i 做 OLS 回归：

$$y_i = x_i\beta + \varepsilon_i$$

这里 β 包括被估参数的一组向量，ε_i 是服从独立同分布假设的随机误差。在经典线性回归中，假设分布为正态分布。当存在空间（或其他）依赖关系时，ε_i 将不再相互独立，并且导致 β 被低估，从而影响假设检验的正确性。

尽管对空间问题的考虑可以追溯到早期在地图制作和调查方面的尝试，但是现代空间回归是直到最近几十年随着统计知识和计算能力的提高才出现的。本书作者向读者介绍了两种应用最广泛的空间回归模型：空间性定距因变量和空间性误差模型，此外还补充了空间分析中的疑难问题。尽管读者的分析单位可能与书中作者的分析单位并不相同，然而书中大量直观的例子仍然能够为读者提供有益的思考。

廖福挺

前 言

空间观念能够为社会科学研究作出巨大贡献。本书试图填补这一领域的空缺，为社会科学家完备地介绍如何将空间依赖性的分析纳入回归框架，让更多对社会科学中空间应用问题感兴趣的读者能够读懂此书。尽管当前已经有非常详尽的空间统计学调查，但是它们中的大部分都过于深奥，并且都假设读者已经具有丰富的高级统计和概率论知识(Banerjee, Carlin & Gelfand, 2004; Cressie, 1993; Getis & Boots, 1978; Haining, 2003; Ripley, 1981, 1988; Schabenberger & Gotway, 2005)。此外，这些调查中的大部分都是自然科学的主题或者该方面的应用，而并不为社会科学研究者所熟知。我们假定此书读者仅仅了解社会科学研究中广泛应用的经典回归模型，同时他们对数据中可能存在的空间依赖性问题感兴趣。在某些部分，虽然我们会用到矩阵的表示形式，但同时也会用非数学语言对其进行详细解释。我们会用到免费且应用广泛的 *R* 计算平台(R Development Core Team, 2004)来演示如何使用这些方法，同时还提供了一段编码，对例子进行解释。如果读者对 *R* 的了解达到了达尔加德一书(Dalgaard, 2002)所介绍的程度，将非常有助于理解本书的内容。尽管其他的程序和方法也可用于分析空间数据，但是本书中将不予以介绍，不过我们会以附录的形式提供一些简要的细节内容。

本书的面世离不开外界的帮助。首先，我们要感谢家庭成员对于我们紧张工作的体谅。同时我们要感谢西班牙巴塞罗那经济分析研究所主任琼·埃斯特班(Joan Esteban)，她给予我们热情慷慨的欢迎和巨大的支持，帮助我们完成了拖延很久的第一稿。迈克尔·沃德还要感谢统计与社会科学中心主任埃德里安·拉夫特里在他访问期间提供的帮助。同时，沃德还要感谢华盛顿大学艺

术与科学学院前院长和现任迈阿密大学校长大卫·霍奇以及华盛顿大学政治科学系主任史蒂夫·马耶斯基的帮助。克里斯蒂安·格里蒂奇得到了来自国家科学基金会的赞助以及来自加泰罗尼亚政府和加州大学的加斯帕·波多拉的旅行资助。

华盛顿大学、加州大学圣地亚哥分校、埃塞克斯大学等地的前任或现任同仁们引导并激发了我们在依赖性数据上的兴趣。我们同样要感谢约翰·阿尔齐斯特、克里斯·巴基、凯尔·比尔兹利、纳撒尼尔·贝克、罗杰·比万德、曹汛、肖娜·费希尔、詹姆士·勒萨热、林泽民、迈克尔·曼格尔、阿西姆·普拉卡什、安德里亚·鲁杰里、伊德里·萨里希安、迈克尔·锡恩、克里斯多夫·沃德、丹尼尔·沃德、安东·韦斯特维尔德三世和艾里克·韦伯尔斯给予的深刻见解和有益的讨论。迈克尔·沃德要感谢他以前在科罗拉多大学行为科学研究所的两位邻座安德鲁·科比和约翰·奥洛林对他在地理方面兴趣的影响。克里斯蒂安·格里蒂奇感谢基德龙的《国家的世界地图集(*State of the World Atlas*)》一书对他在地理和社会科学方面兴趣的影响。

在20世纪,迈克尔·列维-贝克鼓励我们从事这个项目。他是一个有耐心的人。

迈克尔·沃德、克里斯蒂安·格里蒂奇

2010年12月

第1章 | 导论

交互作用与社会科学

社会科学研究者总是对各种情况下不同行动的中间人(例如:个人、政治团体、群体和国家)之间的交往感兴趣。在很多情况下,个体行动者的行动动力和结果并不完全取决于个人特征,而是取决于社会结构及个人的社会位置,以及个体与个体之间的互动。即便是像流感这么平常的一件事情,也有社会因素的作用在里面,因为流感的传播也需要通过社会交往。比如想要预测一个人是否有可能染上鼻病毒,我们就需要考察最近周围是否有异常发生,同时这个人和感染鼻病毒的人是否接触过。有一些疾病是通过接触传染的,也就是感染者在和其他人的交往当中传播疾病。显然,不同类型的交往会导致不同的疾病。在20世纪70年代末期,艾滋病在美国的传播方式被指认为来自加拿大航空公司的单身服务员(Watt, 2003),但这已经被确认为误传。

奇怪的是,交往的作用以及交往结构在社会科学研究中却几乎完全被忽略。比如说,在投票数这件事上,投票率差异以前都是通过个体特征,诸如教育高低或者对政治行为的重视程度来解释的。然而,社会交往以及个体之间的相互联系是与个人特征同样重要的因素。例如,动员投票的电话平均会使投票率变动6个百分点(±3%)(Imai, 2005)。类似的,个人和教堂、工会等组织有联系,也会增加投票率。贝别克和哈克菲尔德(Baybeck & Huckfeldt, 2002)就发现:即便是在分散化的网络当中,间隔较远的个人之间也更少有频繁的交往。这种研究通常是一些例外,而不是惯例。大部分投票率研究仍然假定所有投票

者的决策是相互独立的。

显然,在流感的例子中,将个体之间看作毫无关联,是一种明显不合逻辑的做法。有一些人可能免疫力更弱,所以更可能在流行病发生时得病。然而,我们不可能在不知道其他人是否染病的情况下,仅仅通过个人特征就预测出个人患流感的几率。又比如,父母通常在收入、睡眠时间、吸烟史上都与子女不相同。然而,一方有某种习惯,另外一方通常也会受到影响。社会关系模型的发展来源于心理学家对分离群体和个体间独立效应和交互效应的兴趣,同时他们试图将这种依赖关系用模型表示出来(参见 Kenny, 1981; Malloy & Kenny, 1986)。

在本书中,我们考察了空间分析的视角如何帮助研究者处理观测值之间的依赖关系及处理空间聚类现象。我们尤其关注两类含有空间因变量的回归模型。第一类是关注含有空间滞后因变量(Spatially Lagged Dependent Variable)的情形。第二类是关注空间性误差(Spatial Error)。我们也意识到:空间性模型其实存在很多有趣的视角,但本书除了介绍空间滞后因变量和空间性误差,将不考察其他问题。尽管这些视角对于社会科学的经验研究大有裨益,然而到目前为止,很多文献仍然没有对此引起重视。这些模型使得我们可以考察一个观测值对其他相近观测值的影响。当然,我们相信这种重要性不仅体现在基本原理方面,最简单的道理其实是因为很多社会现象都是空间性"聚集"的。这些空间上排列的数据,既包括地理上观测到的个体位置,也包括在某个地理区域里的地区性数据。后一种数据类型称之为地区(area)或晶格(lattice)数据,而前者称之为点状(point)数据。在本书中,我们将重点放在地区性数据上,这种数据通常用于处理县、州、省、国家等个体单位[1]。

社会科学中空间聚类现象是非常普遍的。投票的地区聚类问题就被认为在美国人的政治行为上起到了重要作用。政治分化与经济与种族上的分化是联系在一起的。正因为如此,投票率模型才不得不将各种分化的空间聚类效应考虑在内(West, 2005)。相似的例子在比较政治学、社会学和经济学研究当中也可以找到。比如,在有关中央银行不同政策的影响的研究中,就有人曾考察过这些政策选择和中央银行以及银行家们的偏好之间的独立性。一种广泛的看法认为:不管中央银行如何不受国家主管机构的影响,它仍然受到各种各样

地方情况的约束。因此，即便中央银行和国家主管机构之间互不干涉，中央银行的政策相互之间也是独立的吗(Adolph, 2004; Franzese, 1999)？默多克，桑德勒和萨金特(Murdoch, Sandler and Sargent, 1997)研究过20世纪80年代欧洲排放硫化物和一氧化二氮的行为，其自愿和非自愿决定之间的相互影响关系。由于污染者们在空间上是分散的，并不受国界的限制，空间分析技术将有助于强调污染的外溢效应以及约束履行上的相互影响问题。在跨国研究中，不平等和贫困问题被认为是交织在一起的。在越穷的国家，财富和收入分配越不均衡。当前研究也发现贪污通常是贫困产生的结果，同时也可能成为贫困产生的原因。然而，研究也发现：收入不平等可能会增加贪污的程度，甚至大于对贫困的影响。一种更复杂的可能的情况是，财富和贪污在空间上具有聚类效应。空间分析就可以帮助我们解决这个难题。尤和卡格拉姆(You and Khargram, 2005)的当前研究就是按照这种思路进行的。最后一点，组织形式的扩展可能也是遵循相同的方式——比如政策效仿。霍姆斯(Holmes, 2006)就利用空间模型探讨过工会组织的蔓延问题。

简言之，社会科学中有无数研究数据都是按照一定的空间形式组织起来的，不管它们的分析单位是县、市、州、国家还是公司。通常，这些分析单位的特征都是高度聚类的，尤其是空间层面的地区聚类。在很多应用中，合乎逻辑的假设都应该包括观测值之间的相互关联性。在实际操作方面，这些聚类通常都被忽略不计或者被当作一种干扰。忽略这种关联性将会极大地影响我们在研究中建立有意义的推论。空间分析不仅为减少这种代价提供了一种方法，同时空间信息将有助于揭示社会过程之间是如何联系起来的。下面我们就将介绍在社会科学的一个重要分支中应用这种分析的简单例子——关于民主制度扩散的研究。

世界各国的民主

在引入空间讨论之前，我们先举一个数据中各观测值之间相互不独立的例子。社会科学家们很早就乐于解释为什么有的国家采用民主政治而有的却不

是如此。早期由李普赛特提出的一项颇具影响力的观点认为：民主制度是具有社会必要条件的。其中之一就是较高的平均收入；李普赛特注意到“在更加民主的国家……平均财富也更高”(Lipset，1959：75)。在过去的40多年里这种观点成为比较分析领域的奠基石，它表明平均收入更高的国家更可能建立民主体制。表1.1提供了一个数据简表，上面列出了2002年世界70多个国家的人均国内生产总值(GDP)和民主水平。我们对民主的测量来自POLITY指数，它将国家按照一系列制度标准划分成不同的类别。在这一指数中，－10代表最不民主的社会，10代表最民主的社会。格里蒂奇和沃德(Gleditsch and Ward，1997)提供了建构这种指数的更详细信息。我们在表1.1中将各国按照人均国内生产值(GDP)和民主程度高低排序，以便于找出变量之间的简单关系。正如我们所见，一些富裕国家，比如丹麦，的确是民主国家；同时低收入国家，比如塞拉利昂和朝鲜，就是专制国家。有趣的是，李普赛特曾提出在1959年，澳大利亚、比利时、加拿大、丹麦、爱尔兰、卢森堡、荷兰、新西兰、挪威、瑞典、瑞士、英国和美国构成了欧洲、北美和南美各洲一系列的“稳定民主体制”。而当时不稳定的民主体制和独裁体制包括奥地利、芬兰、法国、前联邦德国、意大利和西班牙。如今这些国家也变成了民主体制，并且基本上稳定了。尽管这些个案同李普赛特的论调是一致的，但是，财富和民主之间是否存在更普遍的强烈相关关系？尽管印度的平均国民收入很低，但它却实行了民主体制；此外尽管印度近年来经历了高速的经济增长，它仍然远远低于经济合作与发展组织(Organization for Economic Cooperation and Development)成员国的水平。同时，中东地区很多专制国家却拥有很高的收入水平，这也是与李普赛特的观点相违背的。为了更一般化地估计这种关系，我们需要一种更加系统性的比较分析。

表1.1　2002年GDP数值

国　家	民主	GDP	国　家	民主	GDP
几内亚	－1	51	伊朗	3	1776
埃塞俄比亚	1	114	马其顿	6	1801
布隆迪	0	120	纳米比亚	6	1870
扎伊尔	0	135	罗马尼亚	8	1941

(续表)

国　家	民主	GDP	国　家	民主	GDP
塞拉利昂	−10	172	阿尔及利亚	−3	2036
厄立特里亚	−7	175	波斯尼亚和黑塞哥维亚	0	2108
马拉维	5	178	泰国	9	2215
伊拉克	−9	181	苏里南	9	2224
几内亚比绍	5	187	危地马拉	8	2257
利比亚	0	194	俄罗斯	7	2279
卢旺达	−4	216	厄瓜多尔	6	2305
莫桑比克	6	217	秘鲁	9	2306
塔吉克斯坦	−1	221	哥伦比亚	7	2342
尼日尔	4	247	约旦	−2	2375
尼泊尔	6	276	斐济	5	2397
布基纳法索	0	315	突尼斯	−4	2436
乍得	−2	317	萨尔瓦多	7	2486
乌干达	−4	320	南非	9	2607
坦桑尼亚	2	330	多米尼加共和国	8	2745
中非	5	333			
⋮	⋮	⋮	⋮	⋮	⋮
土库曼斯坦	−9	1241	加拿大	10	25139
摩洛哥	−6	1300	芬兰	10	26235
刚果	−5	1303	澳大利亚	10	26304
吉布提	2	1313	荷兰	10	27059
白俄罗斯	−7	1359	瑞典	10	27497
			英国	10	27650
斯威士兰	−9	1412	日本	10	31731
阿尔巴尼亚	5	1416	阿联酋	−8	34436
叙利亚	−7	1417	卡塔尔	−10	36611
哈萨克斯坦	−6	1437	丹麦	10	37063
塞尔维亚	7	1573	瑞士	10	39769
埃及	−6	1602	美国	10	40180
缅甸	−7	1729	挪威	10	43895
保加利亚	9	1744	卢森堡	10	54255

注:GDP 值代表人均国内生产总值。本表有删节,全部数据可以在本册书网站上获取。

自李普赛特(Lipset, 1959)和以往其他学者的研究之后,有关民主的实证比较研究中通常都将民主作为人均 GDP 自然对数的线性函数。我们用 POLITY

分数估计一个国家的民主水平，在纳入人均 GDP 之后，将其表示为普通最小二乘法(OLS)回归：

$$\text{POLITYscore} = \beta_0 + \beta_1 \ln \text{GDPpercapita} + \varepsilon$$

该民主水平对于人均 GDP 的线性回归模型的估计结果见表 1.2。人均 GDP 自然对数的正向回归系数表明民主和收入之间是正向关系，但是如果考虑到变量的测量单位，估计得到的影响效果其实是较小的。

表 1.2　将民主视作人均 GDP 对数的线性方程的 OLS 估计

	$\hat{\beta}$	$SE(\hat{\beta})$	*t Value*
截距(Intercept)	−9.69	2.43	−3.99
人均 GDP 对数(ln GDP per capita)	1.69	0.31	5.36
$N = 158$			
Log likelihood ($df = 3$) $= -513.62$			
$F = 28.77$ ($df_1 = 1$, $df_2 = 156$)			

注：估计值来自于 POLITY 项目和世界银行 2002 年数据。

确切地说，此回归模型预测如果一个国家具有乌兹别克斯坦的人均 GDP (在 2002 年为 464 美元)，那么这个国家的民主得分将接近于 1。反过来，如果一个国家的人均 GDP 收入接近乌兹别克斯坦的两倍(1020 美元)，该模型就可以预测相对应的民主得分约为 2。对于大多数分析者来说，在 POLITY 民主得分指数中，得分 1 和 2 是非常相似的。因此，即便收入水平上存在相对较大的差异，对民主水平的预测结果也不会相差太大，尽管人均 GDP 对数的估计系数在统计上是显著的。

图 1.1 表明 OLS 方程对贫穷国家民主水平的预测远远高于它们的实际水平。在贫穷国家中(比如乌兹别克斯坦)，财富对民主潜在影响的估计效应不仅很小(即使人均 GDP 翻倍，对民主的影响作用也是很小的)，而且这种效应还可能是被高估了。对这些残差的任何标准分析几乎都印证了该图给人的第一印象：这些残差看上去并没有“很好地分布”，因为在最高值和最低值附近存在观测值的两个峰值，这表明模型低估或者高估了实际的民主水平。

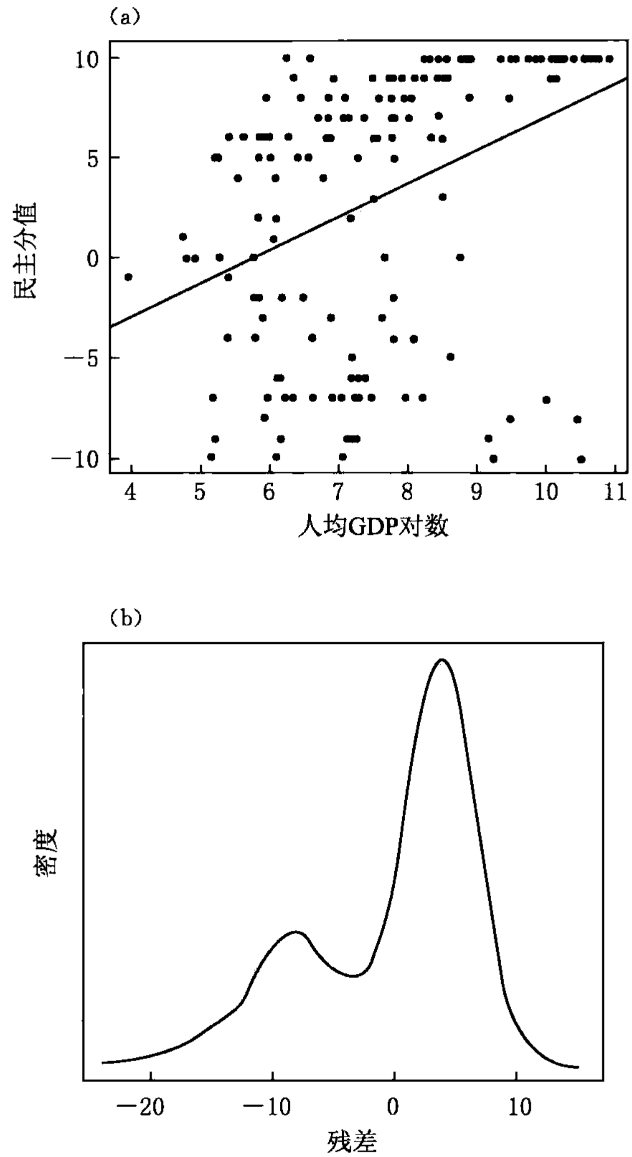


图 1.1 (a)民主和人均 GDP 自然对数的散点图以及回归直线;(b)OLS 残差密度图。它们表明分布较低端存在独裁体制的"聚类"效应和分布较高端存在民主体制的"聚类"效应。

图 1.1 同时也表明围绕估计回归直线,或者说总体趋势上存在大量和呈规律的变化。但是这些残差的排列方式是由于观测值之间的相互关联造成的吗?图 1.1 的(b)图有力地表明残差并不呈正态分布,并且也不是一种单一模式;而是在值−10 附近有一组负值,同时在 5 附近聚集着一组正值。因此,这个例子明确表明表 1.2 报告的 OLS 回归残差是有问题的,进一步回归得到的估计系数

是否可信，也成为疑问。这些残差表明：该模型没有很好地把握民主和经济产出之间的关系，部分原因可能是因为数据之间的关联，即相似值之间的聚集作用。这可能是因为国家之间的相互影响所致。

空间依赖关系介绍

对于以上结果，一种可能的解释是：除了各个国家自身的特性以外，一个国家是否希望建立民主制度和周围国家是否已经建立民主制度是相互关联的。在冷战期间，苏联的干预就使得社会主义制度在很多东欧国家推行。此外，很多拉美国家发生的民主转型似乎也受到了其他国家社会进程的影响（参见 Gleditsch & Ward, 2007）。如果将表 1：1 的数据按照字母顺序排列，我们将很难判断是否存在一些相似体制国家所构成的区域，不同于我们从人均 GDP 中得到的预期类型。即便是按照国家的重要特征排列以后再进行比较，我们也需要经过仔细分析才能识别各种不同类型。

在很多情况下，检验可能存在的空间（以及类似空间的）聚集效应都是非常重要的，它将有助于我们发现是否因为表面的无关联而忽略了内在的社会互动关系。潜在且未观测到的聚集效应，可能会影响我们对于已有模型的理解，从而影响我们真正发现背后的实际过程。在讨论空间相关性之前，我们会先解释为什么这样做很重要。

即便分析者们只是想要比较均值和建立经典统计检验，例如均值差检验，如果数据中存在空间相关关系，那么这种做法也是存在问题的。假设有一个针对变量 y 的单样本的 t 检验，如下：

$$t = \frac{1/n\sum_{i=1}^{n} y_i}{\sigma/\sqrt{n}}$$

如果邻近的观测值之间在时间上或者空间上相关（一阶序列相关），那么对于正向序列相关的值，它们的实际标准误将会偏大（对于负向值将会偏小）。研究者们对于不同时间点上观测值之间的序列相关问题一直比较谨慎，但是他们常常

忽略的是，即便是在同一时间点上，不同观测值之间也可能出现同样的序列相关问题。通过方差的非调整估计得到的 t 值将比真实值更大。这将增大第一类错误(Type I)发生的可能性，即便是在空间自相关作用很小、观测值很多的情况下，也不例外。

简言之，由于观测值之间的空间序列相关(或其他原因)，通过经典检验接受的假设结果将是有偏的，即便在检验结果不真实的情况下也是如此。假设数据在空间上是关联的，例如这种关联与观测值之间的距离成反比，ρ 代表一阶序列相关的空间相关系数。这种相关测量了相邻值之间在一些测量属性上的相似程度。这种相关导致均值的真实标准误将近似于：

$$\sigma_{\bar{y}} \approx \sqrt{\frac{1+\rho}{1-\rho}}\,\frac{\sigma}{\sqrt{n}}$$

一种简单的理解空间相关的影响的方法，是假设一个变量 y 有 n 个观测值：y_1，y_2，…，y_{n-1}，y_n。在很多情况下，我们都认为这些观测值之间相互独立，并且服从同一分布，一般是具有未知均值 μ 和方差 σ^2 的正态分布。对于 μ 的一般估计值为：

$$\bar{y} = \sum_{i=1}^{n} y_i / n$$

由于假定观测值来自于正态分布，那么统计推断将基于 y 和 σ。95%的置信区间为 $\bar{y} \pm 1.96\sigma/\sqrt{n}$。如果 y_i 之间存在空间相关，也就是观测值 y_i 和 y_j 空间上隔得越近相似性越大，那么如同克里斯(Cressie，1993：14)指出的一样，对于取值为正的 ρ，其协方差将为：

$$\mathrm{cov}(y_i, y_j) = \sigma^2 \times \rho^{|i-j|}$$

其方差为：

$$\mathrm{var}(\bar{y}) = n^{-2}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{cov}(y_i, y_j)\right\}$$

展开即为：

$$\mathrm{var}(\bar{y}) = \left\{\frac{\sigma^2}{n}\right\}\left[1 + 2\left\{\frac{\rho}{1-\rho}\right\}\left\{1-\frac{1}{n}\right\} - 2\left\{\frac{\rho}{1-\rho}\right\}^2 \frac{1-\rho^{n-1}}{n}\right]$$

其中的因子：

$$\left[1+2\left\{\frac{\rho}{1-\rho}\right\}\left\{1-\frac{1}{n}\right\}-2\left\{\frac{\rho}{1-\rho}\right\}^{2}\frac{1-\rho^{n-1}}{n}\right]$$

实质上是根据空间相关程度对观测值的数量打了折扣，并且这种作用不会随着样本量增大而消失。如果 $n=10$ 和 $\rho=0.26$（按照克里斯的例子），那么这种折扣效应约为 40%：10 个空间相关的观测值的精度相当于 6 个独立观测值。换句话说，这也表明当观测值之间存在空间正向相关的时候，忽略这种相关所得到的置信区间将远远窄于真实的情况。一般来说，忽视空间依赖性，将可能导致对于数据真实方差的低估。因此，对于一个包含 158 个 GDP 观测值的样本，在正态分布假设下，其 95%的置信区间为 $(1.96\times\sigma)/\sqrt{n}$，但是如果存在 0.65 的空间相关——上面例子中 GDP 的真实 $\hat{\rho}$ 值——那么正确的置信区间将会接近于 4.22 而不是 1.96，几乎是原来的两倍。在民主发展程度的例子中，$\hat{\rho}$ 为 0.47，那么 95%的置信区间为 $(3.26\times\sigma)/\sqrt{n}$，几乎增大了 70%[2]。

如果空间相关具有不同的形式，那么就需要进行不同的具体调整，但是总的原则是如果空间相关为正，那么样本均值的精确性将降低。这通常将导致我们拒绝一个实际为真的零假设。此外，如果数据在空间上具有依赖关系（或者相互关联），那么基于独立同分布（iid）的假设进行的统计检验将变得很不可靠。斯卡本伯格和果特威（Schabenberger and Gotway，2005）详细叙述了在不同样本量和不同程度的自相关情况下最小二乘估计的过度变化。对于 $\rho>0$，这种过度变化随着 n 的增大而增大，比如当 $\rho=0.9$、样本量接近 50 的时候，这种过度变化将接近于 14.0。这里最重要的一点是，空间相关的数据将使基于 iid 假设的统计检验出现严重的问题，这也使得研究者因为标准检验低估了数据的变化而拒绝零假设。

将地图作为可视化数据

人们很擅长发现模式，即便是在没有模式存在的情况下，也是如此。通常

这也就是统计的作用。然而，在具有探索性和启发性的模型中，尽可能多地了解数据是非常有用的。包含丰富信息的密集表格是传递大量信息的重要方式，但是这种做法比较慢。图形展示作为一种辅助方法，将有利于从视觉上很快发现模式的存在。然而，重要的是，这种图形技术应该用于提供研究现象的可能解释上。当前的研究已经阐述了仔细展示证据和定量材料的重要性，并且提供了黄金准则(Cleveland，1993；Tufte，1990，1992，1997；Wainer，2004)。一个指导性的原则是展示方法应当与已有的解释之间存在密切关系。

一项有关伦敦19世纪中叶霍乱传播的经典研究就提供了一个基于地理解释的范例。该研究最早由约翰·斯诺(John Snow)提出，后被蒂夫特(Tufte，1997)推广，最近被约翰逊(Johnson，2006)进一步完善。斯诺证明：1854年夏，伦敦霍乱爆发的原因在于家庭办公一族(以及其他人)的饮水来自于布罗德大街的水泵，而这些水受到来自于霍乱受难者墓场的污染。因此，靠近布罗德大街的水井就成为感染霍乱的潜在危险因素，同时这一研究也对否定霍乱的空气传播论起到了重要作用。斯诺的伦敦地图也成为利用空间相关展现因果联系的重要例子。图1.2提供了家庭办公地区的经典地图，从中可以看出，霍乱所导致的死亡大多聚集在布罗德大街的水泵附近。

在地图上标注阴影也是展现包含地理因素的发生过程的一种重要方式。我们的例子表明：邻近的国家之间会存在相互反馈，这将影响它们的政治体制和经济财富。图1.3的世界地图将158个国家在2002年的民主发展程度用阴影标示出来。这个地图告诉我们民主体制在相邻国家之间有聚集作用，同时在世界不同区域内的专制国家也是相互聚集的。在该地图中阴影越深，表示民主发展程度越高。

只有那些具有最高民主得分的国家才被标记为黑色；例如法国被标记为第二深的阴影，这是因为它的民主得分为9，这也反映了它的总统和国民议会之间相对独立。图1.3表明：民主制度和专制制度都存在很强的地理聚集效应。总的来说，大多数民主国家都位于西欧和美洲或者澳洲和大洋洲沿岸，而很多专制国家都位于非洲、中东和亚洲。



图 1.2 约翰·斯诺关于 1854 年夏伦敦家庭办公(Soho)区的霍乱死亡地图



图 1.3 高民主水平国家用深色阴影表示，它们在世界范围内具有聚集趋势。类似的，更专制的国家在地理上也具有聚集趋势。民主水平用 POLITY IV 指标测量。

当然也存在例外,比如白俄罗斯,尽管周围大都是民主邻国,但它却依然保持着专制体制。相比之下,印度是一个民主国家,尽管周围邻国大多是非民主体制。

在拉丁美洲,尽管国家之间在人均 GDP 上存在巨大差异,但大部分国家在 2002 年都是民主国家。相对而言,尽管中东国家的人均 GDP 水平普遍比世界平均水平高,但其中大部分都是专制体制。事实上,地图上的这些特征告诉我们:民主和人均 GDP 都是在空间上聚集的。在很多情况下,可视化和地图化都有利于揭示数据结构,而这种结构在表格形式的数据中是不容易发现的。

图 1.4 展示了 2002 年人均 GDP(取对数)的聚集情况。富裕的国家用更深的阴影标出,而更贫困的国家用更浅的阴影。图 1.4 也反映了很强的聚集效应。北美和西欧都是富裕国家的聚集地区,而非洲则表明通常贫困国家的邻国也处于贫困之中。当然也存在例外。比如日本和澳大利亚总体来说都比它们的邻国要富裕很多。

将数据地图化将极大地帮助我们解释空间数据,但是好的展示也应该包括经验或者理论的解释。

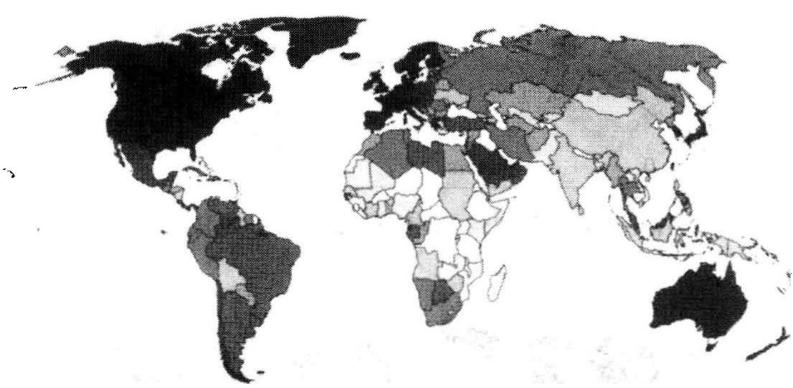


图 1.4 具有更高人均 GDP 水平的国家用更深的灰色表示

空间依赖性和相关性测量

可惜,正如研究者可能忽略数据矩阵中某种模式一样,他们可能在没有结构存在的情况下也找出了某种结构。正因为如此,我们有必要寻找更正式的方式,来衡量观测值是否存在空间聚集,或者相互之间是否通过某种联系关联在

一起。我们将在下一节转向正式的探索方法。

然而，解释这些关联需要我们已知哪些观测值有可能相互联系。对于包括n个元素的一组数据集，每个观测值i可能会和剩下的$(n-i)$个可能的元素相关，但是，在实际中，我们通常假设某些相关关系或者联系将比其他的更重要。我们感兴趣的元素之间的网络或者结构通常必须在分析其他元素之间的关联关系之前就被明确指定。这里我们用到的技术通常起始于关联观测值之间关系的一幅图或者一个列表L。基于种种原因，符合实际的做法是利用矩阵来表示观测值之间的关联性。比如，我们定义一个二元矩阵C来指定个体观测值之间的联系。如果两个观测值i和j被认为相互关联，那么输入值$C_{ij}=1$，反之则$C_{ij}=0$。

测量空间关联和相关的基本思路类似于向量内积，根据休伯特（Hubert），戈利（Golledge）和科斯坦索（Constanzo，1981）的看法，这表示将一个空间接近性的测量与另外一个在某些特殊属性[2]上的相似性的测量交叉相乘。令S_{ij}为两个观测值i和j的空间接近性的测量，同时令U_{ij}表示所关注的某个潜在变量的相似性。向量内积的统计量的一般形式为：

$$\sum_{i=1}^{n}\sum_{j=1}^{n}S_{ij}U_{ij}\text{，}\ \forall i\neq j$$

如果相似性U_{ij}被定义为某潜在变量均值正态化后得到的内积（mean-normalized cross-product），比如$[(y_i-\bar{y})(y_j-\bar{y})]$，那么经过适当的比例调整，再将所有观测值的这个量加总，就可以得到一个空间相关性的测量，称之为莫兰（Moran）$\mathcal{I}$统计量。如果U_{ij}被定义为差值的平方，比如(y_i-y_j)[3]，得到的统计量就被称为吉尔里（Geary's）$\mathcal{C}$。在本书中我们主要关注莫兰（Moran）$\mathcal{I}$。[4]

例如，在测量民主的例子中，空间相关性将涉及测量国家之间一些空间指标上的邻近程度（比如国家之间是否在200千米内接壤）和每对国家在民主得分上的相似性。这些统计指标将有助于发现或者探索空间模式。可能它们最有用的地方在于探索诊断那些原本模型中没有考虑到的空间模式的残差。

估算这些相关性的首要目标是确定数据之间的相互关联作用。这要求我们给出观测值之间如何关联的列表。尽管这一步非常重要，但除此处之外，我们将不会进行过多的阐述。数据之间的连接可以通过物理上的距离来建立，比

如首都之间的距离。然而,其他的传输途径比如公路、铁路、水路以及航运等交通网络的密度在某些情况下可能是一个更好的连接指标。类似的,除了首都城市之间的距离,学者们也曾使用两个相邻国家之间的边界长度,作为它们交往机会的测量。在格里蒂奇和沃德(Gleditsch and Ward, 2001)的研究中开发了一个数据库来记录世界上所有国家之间的最短距离。这里我们将使用该数据,如果国家之间的最小距离小于200公里,那么认为它们是邻国。[5]

表1.3按照两种方式列出了这些数据的一个子集,首先按照列表形式,然后按照矩阵形式。很多电脑程序将大的矩阵按照列表的形式排列,因为这样仅仅将非零的元素保存在记忆体中,可以更有效地存储信息。实际上,对于小的子集,列表存储方式更便于存储数据,且更有利于推导空间特征。

表1.3 欧洲国家子集的连接矩阵

条列清单式

国家	连接线
丹麦	德国,挪威,瑞典
芬兰	挪威,瑞典
法国	德国,意大利,英国
德国	丹麦,法国,意大利,瑞典
意大利	法国,德国
挪威	丹麦,芬兰,瑞典
瑞典	丹麦,芬兰,德国,挪威
英国	法国

连接矩阵格式

	丹麦	芬兰	法国	德国	意大利	挪威	瑞典	英国
丹麦	0	0	0	1	0	1	1	0
芬兰	0	0	0	0	0	1	1	0
法国	0	0	0	1	1	0	0	1
德国	1	0	1	0	1	0	1	0
意大利	0	0	1	1	0	0	0	0
挪威	1	1	0	0	0	0	1	0
瑞典	1	1	0	1	0	1	0	0
英国	0	0	1	0	0	0	0	0

然而,每个表都能很容易地被转换成一个方形矩阵,该矩阵描绘了沿行和列排列的观测值以及矩阵内部的关联。矩阵形式同样有利于明确变量定义或反映

空间结构和变化的测量。表 1.3 的第一部分表示一组列表形式的连接数据；第二部分是相应表示二元连接的矩阵 C。

这些数据也可以通过简单网络图来展示，如图 1.5。这种图非常具有启发性，但是一旦节点数量很多，它们就会变得难以阅读。图 1.5(b)就展示了 158

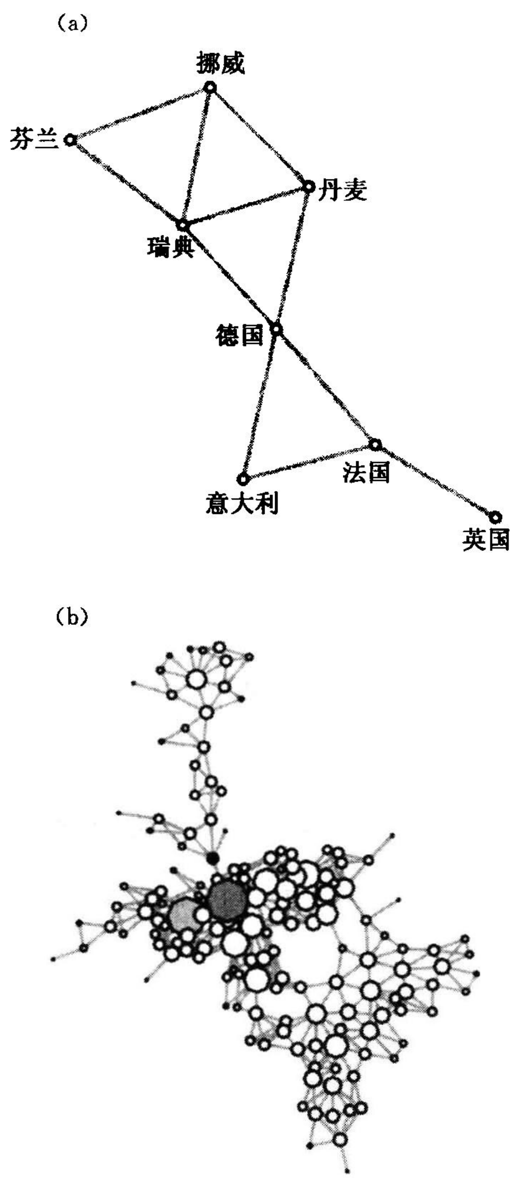


图 1.5 表 1.3 中 158 个国家数据的简单网络：(a)8 个欧洲国家之间的关联；(b)158 个国家之间的关联

注：在(b)组中美国是黑色而俄罗斯为灰色阴影。节点大小与 200 千米以内国家的数量成比例。

个国家的网络图的拥挤情况。然而，这种视觉网络表现法可能会对某些数据，尤其是小数据的情况非常有用。

一旦构造一个观测值之间连接的可能网络，用列表 L 表示或者用连接矩阵 C 表示，我们就可以试图找出所关注的某个变量的取值（这里用 y 来表示）是否与相互连接或者相邻的观测值相似。一种可行的方法是观察两个相连的观测值 i 和 j 是否相似，比如：判断 i 观测值的高低是否和 j 观测值的高低共变。通常 i 会和很多观测值相连，除非它和很多邻近观测值都相似，否则空间聚集效应也不会存在。为了整合相互连接的观测值的有关信息，我们通常假定所有的邻近观测都具有相同权重，并且每一个的权重都将是 1 比上总的连接数量的比例。计算空间滞后（Spatial lag）的主要目标，是得出周围区域的均值。美国周围邻国的平均民主得分是多少呢？加纳邻国的平均人均 GDP 是多少呢？这些邻国的均值与每个国家自身的民主得分或者人均 GDP 有关吗？我们提供了一个探索性的统计量，来测量空间相关性。研究者可以以同样的方式生成独立变量之间的相关矩阵，这种空间相关可能会为观测数据提供探索性的信息。

令 y_i^s 表示 y 的所有相关观测值的均值或者平均数，或者叫 y 在空间上的“滞后（lag）”。矩阵表达方式将有助于发现基于 y 所建立的空间滞后 y_i^s 以及连接矩阵 C。我们可以构造一个行标准化的连接权重矩阵 W，该二元连接矩阵 C 中将每一个行向量 c_i 除以总的连接数 $\sum c_i$，使每一行加起来为 1。表 1.4 中给出了一个例子。

表 1.4　含有 8 个欧洲国家的数据子集的行标准化连接矩阵

	丹　麦	芬　兰	法　国	德　国	意大利	挪　威	瑞　典	英　国
丹　麦	0	0	0	1/3	0	1/3	1/3	0
芬　兰	0	0	0	0	0	1/2	1/2	0
法　国	0	0	0	1/3	1/3	0	0	1/3
德　国	1/4	0	1/4	0	1/4	0	1/4	0
意大利	0	0	1/2	1/2	0	0	0	0
挪　威	1/3	1/3	0	0	0	0	1/3	0
瑞　典	1/4	1/4	0	1/4	0	1/4	0	0
英　国	0	0	1	0	0	0	0	0

注：连接存在于国境线距离小于 200 千米的国家之间。

在这种情况下，标量 $y_i^s = c_i y$（通过相加）计算了元素 i 所有相邻观测值的平均值或均值。这通常称为空间滞后。$y^s = Wy$ 这组关系告诉我们每一个 y_i^s 和其他国家的 y 以及连接权重 w_i 之间的关系。表 1.5 列出了民主变量的 10 个最大的正向和负向空间滞后。例如，巴林的民主得分为－8，但是它周围的所有邻国都是具有最大负向民主得分－10。另一方面，爱尔兰和葡萄牙具有最高民主得分，而它们的邻国也一样。

表 1.5　最大和最小空间滞后的 10 个国家

国　家	民主	空间滞后	国　家	民主	空间滞后
最大负向空间滞后			最大正向空间滞后		
巴　林	－8	－10	卢森堡	10	9.8
塔吉克斯坦	－1	－7.1	瑞　士	10	9.8
阿　曼	－9	－6.7	英　国	10	9.8
吉尔吉斯斯坦	－3	－6.6	比利时	10	9.8
阿联酋	－8	－6.5	荷　兰	10	9.8
乌兹别克斯坦	－9	－6	加拿大	10	10
卡塔尔	－10	－5.8	斐　济	5	10
也　门	－2	－5.5	法　国	9	10
科威特	－7	－5.3	爱尔兰	10	10
以色列	10	－5	葡萄牙	10	10

注：对应于每个国家，列出的是它们相应的民主得分和空间滞后的民主得分。

接近性测量

对很多社会科学家来说，空间分析中最重要的一步是测量个体间的接近性。在社会环境中，距离是什么呢？比如，尽管物理学家们可以用严格的地理或者欧几里得距离测量树和树之间的距离，但测量距离在社会科学分析中却要复杂得多。例如：美国和墨西哥之间隔多远呢？如果我们使用严格的连续性测量，这两个国家是最理想化情况下的邻国，因为它们具有共同的边境线。但是加拿大和美国也具有共同的边境线。这表明它和美国的距离与前者一样大吗？从华盛顿特区到墨西哥城的直线距离是 3000 千米，但是从华盛顿特区到渥太华

的距离为 700 千米。我们也可以使用国家之间边境线的长短或者最大的 10 个人口中心的平均距离来测量国家之间的距离。图 1.6 给出了这两种不同识别方法的区别。一些国家版图的中心(空心圆点)和它们的首都(黑色圆点)之间的距离非常远,但是,在一些小国就不会出现这种问题。中国、加拿大、俄罗斯、澳大利亚和美国都属于两种圆点之间距离较远的类型。相对而言,在朝鲜和韩国,版图中心和首都之间的距离就非常小。

在实际应用中,另外一个重要的问题是如何处理缺失的空间数据。插补方法(Imputation)就是一种处理方法,当然也有其他办法(Griffith, 2003)。真正的问题在于社会科学的数据通常存在缺失问题,而且很少是随机缺失的。在不是空间数据的情况下,这种缺失可以通过标准方法处理——利用插补方法,或者更常用的是将删除有缺失信息的观测值。然而在空间数据结构中,这些缺失可能会在空间图上产生一些"洞(holes)",从而使我们不能准确完整地表现空间上的接近性。在空间结构上,另一些可能出现的问题是,某些观测值不与其他观测值相连。例如,新西兰在 200 千米之内就没有其他独立的国家。有两种避免出现这种情况的常用的办法。通常我们会从分析中删去岛国,这是因为在很大程度上它们没有连接对象,从而也不会影响研究中其他观测值的空间过程。更显而易见的原因是,删除它们之后,可以排除空间加权矩阵中的奇异阵情况(也就是行或者列完全由 0 组成)。第二种方法是选取岛屿最邻近或者最可能的邻国,比如将澳大利亚作为新西兰的邻国连接,即便其他观测值都选取 200 千米,作为它们之间是否有连接的标准。更一般的情况是,可以对所有的个体选取最近的 k 个邻国距离。

图 1.6 可以通过下列的 R 命令生成。

```
# Set working directory
dd <- c("C:...")
setwd(dd)

# Plotting map with centroids and capitals
```

```
# Load required libraries
library(RColorBrewer); library(maptools)
library(spdep); library(sp); library(rgdal)

# Read a Robinson projection map from an ESPI shapefile
rob.shp <- read.shape("wg2002worldmap.shp")

# Indicate the id codes for each polygon/country
rob.map <- Map2poly(rob.shp, region.id =
    unique(as.character(rob.shp$att.data$FIPS_CNTRY)))

# Indicate the map projection
tr <- readShapePoly("wg2002worldmap", IDvar = "FIPS_CNTRY",
    proj4string = CRS(" + proj = robin + lon 0 = 0"))

# Extract the relevant variables and exclude missing data
ct <- na.omit(rob.shp$att.data[,c(1, 18:20)]

# Assign relevant variable/column names
colnames(ct) <- c("ID", "x", "y", "City_POP")
ct$x <- as.numeric(as.character(ct$x))
ct$y <- as.numeric(as.character(ct$y))

# Add coordinates
coordinates(ct) <- c("x", "y")
proj4string(ct) <- CRS(" + proj = longlat + datum = WGS84")

# Transform the coordinates to the robinson projection
ct_rb <- spTransform(ct, CRS = CRS(" + proj = robin + lon0 = 0"))

# Replot the map itself without a bounding box
```

```
plot(rob.map, border = "Grey", forcefill = T, xaxt = "n", yaxt = "n",
    bty = "n", lwd = .000000000125, las = 1, ylab = "",
    main = "Centroids and Capitals", xlab = "")

# Add the centroids
points(coordinates(tr), pch = 19, cex = .5, col = "grey")

# Add the capitals
points(coordinates(ct_rb), pch = 19, cex = .5, col = "black")

# Add segments between centroids and capitals
tr_or <- coordinates(tr)
rownames(tr_or) <-
    as.character((attributes(tr) $ data) $ FIP S_CNTRY)
ct_rb_or <- coordinates(ct_rb)
rownames(ct_rb_or) <- as.character(ct_rb $ ID)

# Delete Kiribati (91), as longitude extends across
    international date line
coor_dif <- cbind(tr_or[ -91,], ct_rb_or[rownames(tr_or),][ -91,1])
x1 <- coor_dif[,1]
x2 <- coor_dif[,3]
y1 <- coor_dif[,2]
y2 <- coor_dif[,4]
segments(x1, y1, x2, y2, col = "slategray4")
```

以上我们列举了两种测量距离的基本方法，但是这还停留在表面。对距离的度量，也可以用平均行程时间、每两点之间的移动电话通话数目、两地之间的旅游观光数量，或者其他任何形式的关于距离与交往的测量。例如，国家之间

有大量的贸易活动，就可以被认为在经济上“密切”(Lofdahl, 2002)。格里菲斯(Griffith, 1996)就提供了一些这种测量的想法和操作方法。

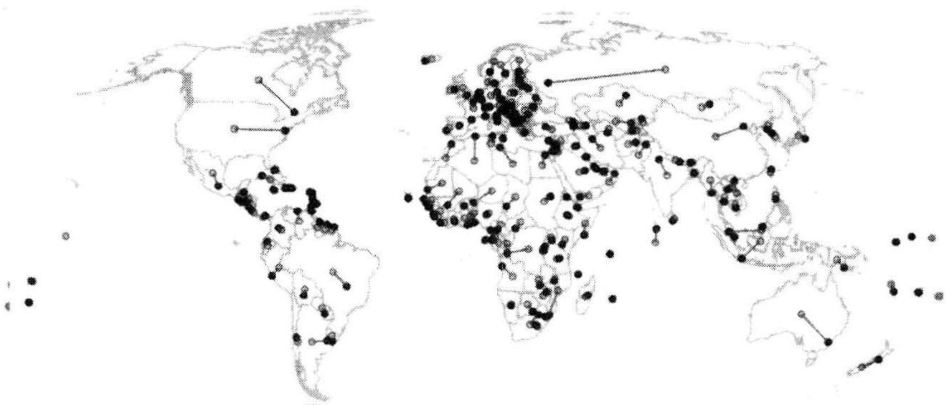


图 1.6 地理中心(空心点)和首都(黑点)之间连线图

我们通过 y 和 y^s 之间的相关程度，来测量国家自身的民主发展水平和它们邻国的民主水平的关联，这看起来是一件合乎常理的事情。某国自身取值与各邻国的加权平均值之间的线性相关被称为莫兰 I 统计量(Moran, 1950s, 1950b)，这个全球相关(global correlation)包括一个观测值和它邻近值之间的所有取值。广义的莫兰 I 统计量用一个加权的成比例的向量内积表示：

$$I=\frac{n\sum_{i}\sum_{j\neq i}w_{ij}(y_i-\bar{y})(y_j-\bar{y})}{(\sum_{i}\sum_{j\neq i}w_{ij})\sum_{i}(y_i-\bar{y})^2}$$

其中 w 表示行标准化的加权矩阵 W，y 是我们所关注的变量。

I 被认为服从正态(渐进)分布，其均值为 $-1/(n-1)$。莫兰 I 的方差可表示为：

$$\mathrm{var}(I)=$$

$$\frac{n^2(n-1)\frac{1}{2}\sum_{i\neq j}(w_{ij}+w_{ji})^2-n(n-1)\sum_{k}(\sum_{j}w_{kj}+\sum_{i}w_{ik})^2-2(\sum_{i\neq j}w_{ij})^2}{(n+1)(n-1)^2(\sum_{i\neq j}w_{ij})^2}$$

如果将所关注的变量标准化为 z_i，莫兰 I 可以简化为：

$$I=\frac{1}{2}\sum_{ij}c_{ij}z_iz_j\quad\forall i\neq j$$

莫兰 I 统计量通常通过建立包含均值和方差信息的 z 值来检验空间相关性。

莫兰 I 并没有一个真正固定的度量标准，但是它的期望值是 $-1/(n-1)$，而不是 0。然而，莫兰 I 统计量可以用图形来解释，帮助我们理解个体之间的不同空间相关程度如何导致该统计量的取值变化。比如一个 $\tilde{y}$ 和其邻近值平均数 $\tilde{y}^s$ 的散点图[这里我们使用标准化的 $\tilde{y} = (y-\bar{y})/sd(y)$ 使得均值为 0，标准差为 1]。在这个图中，观测值在四个象限里围绕 $\tilde{y}$ 和 $\tilde{y}^s$ 的均值分布，这反映了变量 y 的空间相关关系。如果 y 没有空间聚集关系或者存在相关关系，y^s 的值将不会随着 y 的变化而发生系统性的变化。然而，如果存在正向空间相关，个体观测值的取值高于或者低于 y 的均值将在 y^s 上（或者说在邻近国家中）相应地反映了高或低①的趋势。大部分的点将落在第一和第三两个象限，在这些位置，个体和它们的邻近值是相似的，而位于第二或第四象限的观测值将会比较少。如果我们对散点图画出它们相应的回归直线，它的斜率就是针对原始变量 y 和连接列表或者矩阵 C 之间的莫兰 I 相关系数。

图 1.7 以一种固定格式说明了莫兰 I 统计量，并解释了变量及其一阶空间滞后量组成的散点图。回归直线的斜率表示数据中空间相关性的平均数；也就是莫兰 I 统计量。

莫兰 I 通过调整 y 的变化以及每个观测值相邻点的数量，比较了 i 的所有邻近点和均值之间偏差的关系。莫兰 I 值越高，表明地理上的聚集作用越强；也就是说，邻近取值的相似性越大。这个统计量测量了一个观测值和它的邻近点之间的平均相关关系。图 1.7 就解释了这个基本概念。空间滞后情况（某点的邻近点的平均值）用纵轴表示，而横轴表示每个观测值的取值，标准化以后使得其均值为 0、方差为 1。方框表示 ± 2，大部分的观测值都落在这个边界内（注意：在这里 $2\sigma = 2$，因为变量经过了标准化）。那些落入阴影方框的观测值表明其拥有同质性很强的邻国。那些落在高于（0，0）和（2，2）阴影区域的观测值表明其取值高于均值，并且它们的邻国的平均取值也高于均值。类似的，落在（0，0）和（−2，−2）②之间阴影区域内的观测值低于均值，同时它们的邻近点也如此。

---

① 原书此处为低或高。应该为笔误。——译者注

② 原书这里为（−2，2），根据上下文应该是一处笔误。——译者注

这个散点图的对角线区域有很多观测值,这反映了相似值的聚集效应。图中也有一些点所代表的观测值自身低于观测变量的平均值,但是,平均而言,其邻近的点却远高于变量的均值。在散点图的左上角部分的一个点就是这种情况,它可以被看作是一个被取值很高的邻近点包围的、而自身取值很低的孤立点。针对这些标准化后的观测值的 OLS 回归,可以生成一个概括观测值及其邻近点之间取值关系的测量值。举个例子,如果我们关注的变量为犯罪率,并收集了休斯敦市每个管辖区的数据,在图的右上部分将是具有高犯罪率的管辖区,其周围也是高犯罪率的地区。类似的,左下角代表自身犯罪率低而周围犯罪率也低的地区。穿过这些标准化点的回归直线的斜率就是莫兰 I 统计量。

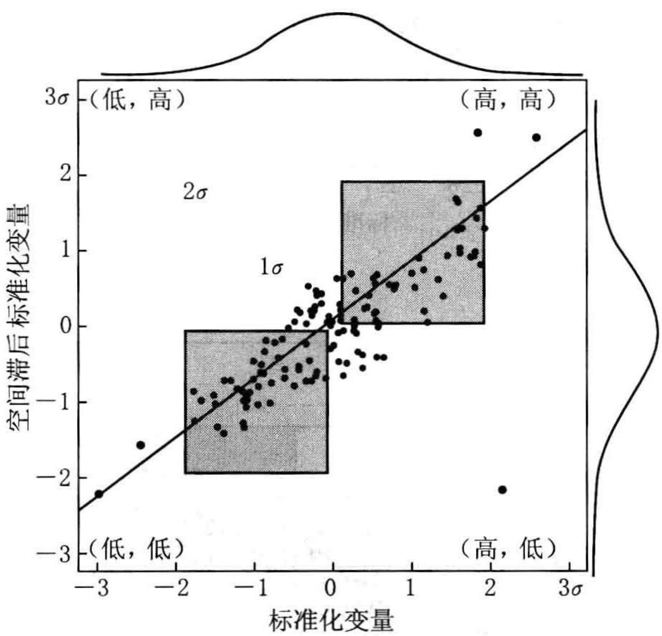


图 1.7 变量和其空间滞后的散点图

注:变量经过标准化使得其均值为 0,方差为 1。同质性的邻近观测值聚集群用阴影部分标示。图中也给出 OLS 回归直线。

针对莫兰 I 系数的统计检验,需要加入其他假设,因为在简单概率检验的经典框架中,需要用到一阶和二阶矩(均值和方差)。在假设检验框架中,经常被忽略的重要一步,是明确零假设。对空间模型而言,由于在不同的空间变量中可能存在各种迥异的模式,因此零假设就显得很不明显了。比如,空间模式是

正态分布的吗？还是随机分布的？如果是随机分布的，那么在空间中是完全随机的吗？一般来说，在当前文献中，有两种解决办法，但它们在一定程度上都是事先设定的(ad hoc)。第一种方法假设数据是正态分布的。克里夫和奥德

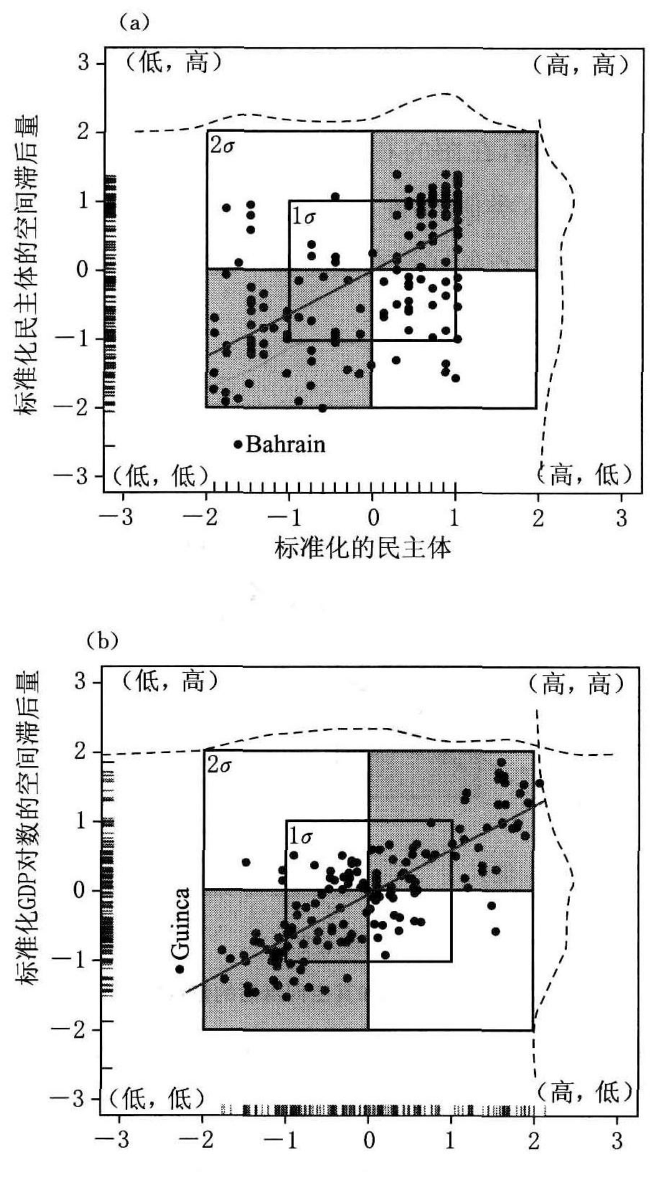


图 1.8　标准化变量与其空间滞后图

(a)民主化，莫兰 I = 0.64；(b)人均 GDP，莫兰 I = 0.65

(Cliff and Ord, 1971)计算出这种情况下I的方差。尽管很多研究表明假设莫兰I为正态分布通常是错误的(Boots and Tiefelsdorf, 2000; Tiefelsdorf, 1972),但是大多数软件和应用文章仍然使用正态性假设[6]。第二种方法是利用蒙特卡洛模拟(Monte Carlo simulation),对连接矩阵随机出足够多次的行列变换,从而得到一个随机化的零模型。大多数统计软件都会提供这两种主要方法供选择,它们通常(当然也不总是)会得到相似的结果。

图1.8(a)给出了民主得分标准化后的散点图。该图将$\pm 2\sigma$用方框标出,为了便于发现哪些观测值过度异常。右上象限表示得分较高并且周围也是较高得分的观测值。位于左下象限的点,表示得分较低并且周围也是相似较低得分的个案,巴林就是其中一个极端的例子。在图1.8中,非对角线上的点表示这些国家和它们邻国的民主水平差别很大。正如图中所示,对于专制国家这样的个案是非常少的(最例外的情况是白俄罗斯),而对于民主国家这样的个案更是少之又少。图中也画出了回归直线,它的斜率就表示民主的莫兰I统计量(为0.64),这个值比该统计量在这个例子中的期望值要大得多(−1/158)。图1.8(b)表示人均GDP。几内亚位于左下角,已经超出了2σ的方框范围;卢森堡在右上角也超出了该方框区域。

我们可以根据表1.2中OLS估计出的残差的莫兰I来考察残差的变化是否表现出空间聚集[7]。这仅仅是一个探索性的考察,在原始数据中使用莫兰I的作用差不多。利用R软件可以很容易地完成,定义一个回归目标ols1.fit并取列表nblist中200千米边界范围内的其他国家作为其"邻国"。

```
source("chapter1data.R")
ols1.fit <- glm(democracy ~ log(gdp.2002/population), data = sldv)
library(spdep)          #Load spdep library for moran.test()
moran.test(resid(ols1.fit),nb2listw(nblist))
lm.morantest(ols1.fit,nb2listw(nblist))
```

根据 OLS 残差计算出的莫兰 I 统计量为 0.40，方差为 0.0028。相应的标准分为 7.77，它比 −1/158 要大得多，同时对应的 p 值≈0。这表明基于观测值之间相互独立的假设得到的 OLS 结果受到因变量和自变量的空间聚类的强烈影响。因此，这可能误导了我们在统计上和实际上推断民主和社会财富（由人均 GDP 表示）之间的关系。

估计空间模型

空间分析的一系列步骤简单来说是怎样的呢？

第一，将数据在地图上标示出来，尤其是因变量。这可以在很多种环境中完成，比如数据表插件（spreadsheet plugin）、地图混搭程序（map mashups），以及 GIS 软件包，但是最好的情况应当是能够统计分析的平台。我们将介绍 R 库的使用，尤其是通过 maptools 和 spdep 创建变量分布的简单地图。

第二，同时，判断因变量上是否有明显的空间相关。对于本书中的大多数方法（即不是点过程[point processes]的方法），这是指计算莫兰 I 统计量来估计空间相关的大小。分析者在有的情况下希望通过局部空间相关指标（Local Indicator of Spatial Association, LISA）来考察和绘出每个观测点对空间相关的贡献。本书将不详细讨论这一点。更多有关的讨论和例子，可参见格里蒂奇和沃德（Gleditsch and Ward, 2000），安瑟林（Anselin, 1995），以及奥德和格蒂斯（Ord and Getis, 1995）的文章。

第三，将这些空间滞后变量准确地合并到基本的统计框架中，并且检验得到的残差是否仍然是空间相关。

第四，除了利用正态模型探索方法估计模型拟合程度和预估参数的不确定性程度，我们也需要计算和检验均衡效应（equilibrium impact）。这表明需要梳理出估计得到的空间模型的均衡效应和相互反馈的作用。

下面我们将根据当前的例子来讲解这些步骤。

将数据地图化同时建立空间权重矩阵

我们已经讲解了如何将2002年158个国家民主得分的数据在地图上标示。在这一节当中，我们将讲解如何将OLS回归残差在地图上标示。在前面的图1.3和图1.4中，数据自身在地图上就已经标示出来了。我们同时发现：使用收入对民主进行回归之后，其残差也表现出空间相关性。我们通过定义200千米距离内的"邻国"计算得到莫兰I。如前面提到，该例子中莫兰I为0.4，其方差为0.0028。在一般情况下这种显著结果让我们确信从图1.3和图1.4中观测到的空间模式实质上影响了回归结果，也就是说，它造成了估计和标准误的偏差。图1.9反映了OLS中的残差。

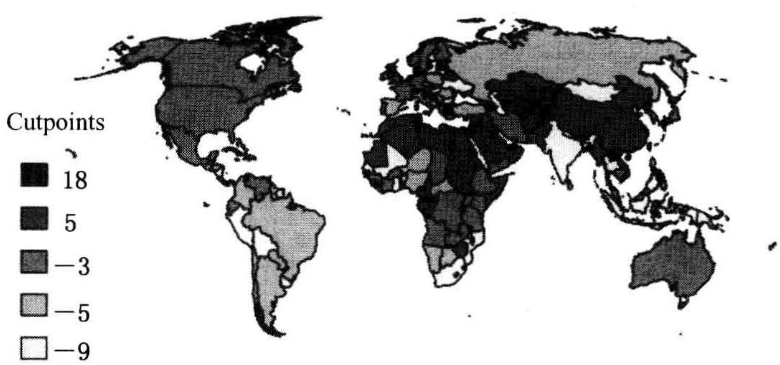


图1.9 在地理上显示OLS回归的残差

寻找空间类型

根据锡恩(Shin, 2001)的研究，我们也可以创建一个锡恩空间散点图(Shin spatial scatterplot)。该图画出了每个输入变量标准化后的值——在这里为残差——和相对应的空间滞后或者是与其相连接的观测值的均值(图1.10)。阴影框表示当某值大于残差均值时，其邻近点也为正值的情况。坐标轴上包括一个"毯图(rug plot)"用于表明变量的分布。外部边缘区域表示对变量本身以及其空间上估计分布的核密度分布(kernel density)。生成此图的代码如下：

```
pdffilename <- c("file name and path")
pdf(file=pdffilename, width=5.0, height=5.0, family="Times")
dem <- (resid(ols1, fit))  #residuals
ds <- (dem-mean(dem))/sqrt(var(dem))  #standardized democracy score
# create spatial lag and standardize it
ds.slag <- as.vector(wmat % * %ds)
ds.slag <- (ds.slag~mean(ds.slag))/sqrt(var(ds.slag))
plot(ds, ds.slag, xlim=c(-3, 3), ylim=c(-3, 3), pch=20, las=1,
          xlab="standardized democracy",
          ylab="spatial lag of standardized democracy")
reg1 <- lm(ds.slag~ds)

# establish a grid
xgrid <- sq(-3, 1.5, length.out=158)
x0 <- list(ds=xgrid)
pred.out<-predict(reg1, x0, interval="confidence")

# put 1 and 2 sigma boxes on plot
lines ( c(-2, -2, +2, +2, -2), c(-2, +2, +2, -2, -2))
lines ( c(-1, -1, +1, +1, -1), c(-1, +1, +1, -1, -1))
lines ( c(-2, +2), c(0, 0))
lines ( c(0, 0), c(-2, +2))

# some text for context
text(-2.5, 3, "(low, high)"; text(2.5, 3, "(high, high)")
text(-2.5, -3, "(low, low)"; text(2.5, -3, "(high, low)")
polygon(x=c(-1, 0, 0, -1), y=c(-1, -1, 0, 0), col="slategray3")
polygon(x=c(0, 1, 1, 0), y=c(0, 0, 1, 1), col="slategray3")
```

```
# plot c.i. region
polygon(x = c(xgrid, rev(xgrid)), y = c(pred.out[, 3],
         rev(pred.out[, 2])), col = "slategray3", border = T)

# put data on plot
points(ds, ds.slag, pch = 20)

# densities
sldensity <- density(ds.slag)
lines(sldensity$y + 2, sldensity$x, lty = 2, col = "slategray4")
ddensity <- density(ds)
lines(ddensity$x), ddensity
$y + 2, lty = 2, col = "slategray4", xlim = c(-2, 2))
points(ds, ds.slag, pch = 20)
lines(xgrid, pred.out[, 1], type = "1", lty = 2, col = "gray80", lwd = 2)

#rugs on two sides
rug(jitter(ds, factor = 2), col = "slategray3")
rug(ds.slag, side = 2, col = "slategray3")

# label some points
text(-2., -2.3, "Oil Exporters", col = "slategray4")
dev.off()
```

接下来我们将考察民主测量上的空间关联性。民主变量的空间滞后，简单来说，就是周围国家的民主水平的平均数。这里邻国民主得分高的国家，其自身的分值也高，同时其邻国专制程度高的国家，其分值也负得越大。我们在图 1.11 中将它们在地图中画出来。该地图显示：位于非洲和亚洲的国家，其邻国都是非民主国家，而欧洲和大部分美洲国家的邻国都为民主国家(Gleditsch, 2002a)。

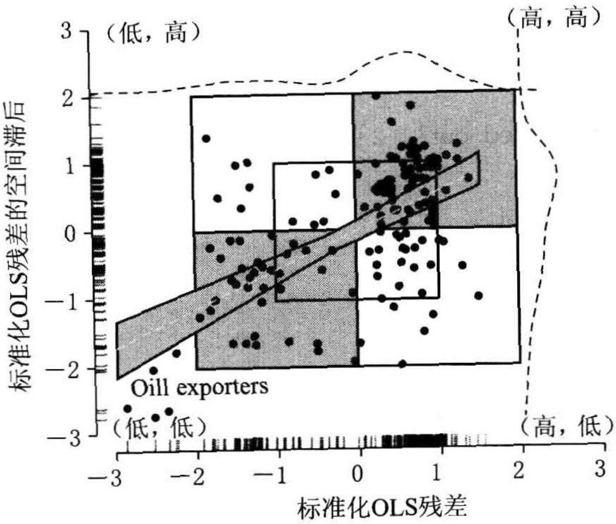


图 1.10　OLS 残差锡恩图(Shin Plot)

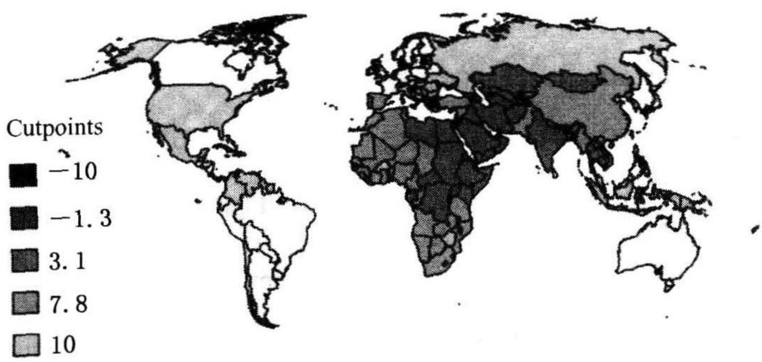


图 1.11　民主的空间滞后性(颜色越深表示空间滞后变量值越大)

其执行代码如下：

```
# sldv2 is the data.frame
# mdd2 is the minimym distance data.frame
nblist <- vector(mode = "list", length = dim(sldv2)[1])
attr(nblist, "region.id") <- sldv2$tla
attr(nblist, "class") <- "nb"
nbnms <- data.frame(sldv2$tla, c(1:dim(sldv2)[1]))
names(nbnms) <- c ("acr", "nm")
```

```
min200 <- mdd2[mdd2$mindist<=200,] # Create an index of the isolates
nodata <- setdiff(sldv2$tla, unique(c(min200$ida, min200$idb)))

# Find neighbors for each row in the sldv for(i in 1:dim(sldv2)[1]){
  temp <- min200[min200$ida==sldv2$tla[i] |
              min200$idb==sldv2$tla[i],]
cty <- unique(c(temp$ida, temp$idb))
cty <- setdiff(cty, sldv2$tla[i])
  nblist[[i]] <- nbnms[match(cty, nbnms$acr), "nm"]
}

# wmat is the row standardized weights matrix
wmat <- matrix(0, ncol=dim(sldv2)[1], nrow=dim(sldv2)[1])
rownames(wmat) <- colnames(wmat) <- sldv2$tla
for (i in 1:dim(min200)[1]){
  wmat[min200$ida[i], min200$idb[i]] <- 1
}

wmat <- wmat/rowSums(wmat)

# Calculate the spatial lag of democracy
Democracy.spatial.lag <- as.vector(wmat %*% sldv2$democracy)
```

除了在地图上标示出一阶民主空间滞后，将每个观测值对总体(Global)莫兰I统计量的贡献在地图上标示出来也是很有用的。这个量就叫做LISA(局部空间相关指标)统计量。这里我们将所有的值标准化，然后将它们在地图中标示出来，如图1.12。局部莫兰统计量的计算来自于奥德和格蒂斯(Ord and Getis, 1995)，安瑟林(Anselin, 1995)以及格蒂斯和奥德(Getis and Ord, 1996)的文章。

该地图表明：从其邻国的民主发展水平来看，哪些国家的情况比较异常。

南非和西非就属于这种情况，印度也是。

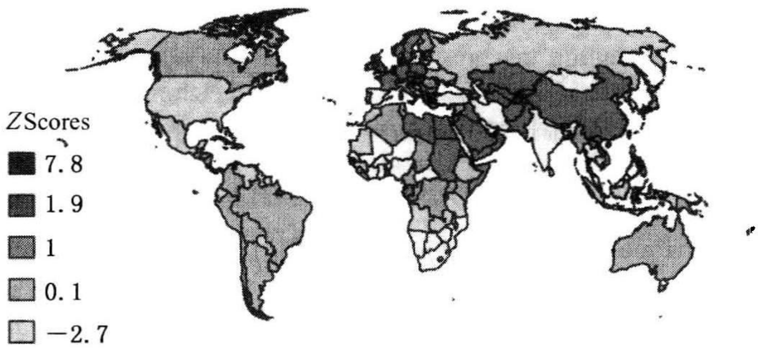


图 1.12　局部莫兰 I 统计量的 z 值

注：用灰色等级指代邻国之间相似性。

小结

首先经过仔细考察数据并将数据视觉化以后，我们基于民主水平和财富（用人均 GDP 的对数来表示）之间关系的线性假设，得到 OLS 回归结果。根据从回归残差中得到的有力证据，我们发现：残差之间表现出空间聚集的关系，这违背了回归中认为个体观测值的误差项应当相互独立的假设。OLS 估计基于个体相互独立的假定，因此，在分析收入和民主之间的关系时，OLS 将不会是一个可靠的模型。更重要的是，模型中假定对于这些独立观测值，民主仅仅受到收入的影响，而忽略了显而易见的地理聚集效应的特征。我们也展示了如何利用地图和简单的统计量来提供有关空间聚集程度及其特性的探索性信息。

即便研究者对回归分析不感兴趣，也可以用社会科学数据发现空间模式。如果仅仅进行简单的均值检验或者用回归方法分析空间排列的数据，而不考虑空间相关关系，都将会带来错误的推论，从而导致错误地拒绝了已有的假设。

通过地图展示相关联的数据，为我们提供了一种判断空间模式是否存在的探索性方法，而空间模式的存在，也会使统计推论变得更加复杂。下面我们将介绍如何估计包含空间滞后因变量的回归模型，这种方法有助于我们明确地将空间依赖关系纳入回归框架中。

第 2 章 | 空间滞后因变量

在本章中，我们将讨论如何将空间依赖性，也就是将包含“空间滞后”因变量的 y 明确地加入回归方程的右边。这种模型有很多不同的名称。安瑟林(Anselin，1998)将其称之为空间自回归(spatial autoregressive)模型，但这个术语可能会引起一些混淆，因为自回归这个词在地理统计文献中表示另外一些完全不同的空间模型。为了简单起见，这里我们称之为空间滞后 y 模型，原因是该模型的主要特征是加入了作为协变量的空间滞后因变量。

当我们确信每个个体 i 的 y 值受到其“周围值”的直接影响的时候，空间滞后 y 模型就很适用。这种影响大于和超过 i 的其他协变量的影响。如果我们相信 y 并没有直接受到周围值的影响，而是因为某些在模型识别中忽略的空间聚集特征同时影响了个体 i 的 y 值及其周围值，这时就需要考虑一种空间相关误差的模型，我们之后会讨论到。在空间滞后 y 模型下，因变量 y 必须被视为连续变量。本书将不讨论更复杂的二分因变量的例子。这是因为更复杂的情况可能无法得到闭合形式的解，同时迭代估计方法也超出了本书的范围(参见 Ward & Gleditsch，2002)。

空间滞后因变量的回归

为了进一步讲解空间滞后 y 模型，我们回到前面世界范围内民主分布的例子。我们已经看到民主的分布展现出空间聚集的效应，也就是如果一个国家周围都是民主水平很高的国家，那么它自身的民主 POLITY 得分也会较高。尽管有一些民主聚集现象可能来自于人均 GDP 的空间聚集，并且它和民主之间也

具有正向相关,但是我们仍然发现:即便是控制了国家的人均 GDP 水平之后,民主的空间聚集效应依旧没有完全消失。在民主对人均 GDP 回归模型中,由于假设误差$\in_i$相互独立,这样利用回归残差就可以检验空间依赖是否存在;也就是说根据$\hat{\in}_i=(\hat{y}_i-y)$,同时利用莫兰 I 相关系数以及识别出的矩阵 C 的连接形式——在这里国家之间如果在 200 千米范围以内,就看作相互连接。在本例中,我们发现了残差间的强空间相关关系。残差的莫兰 I 统计量为 0.40,相应的 z 值近似于 8[8]。这个结果远远大于零假设(即空间独立性)为真的情况下的值。换句话说,这表明国家民主发展水平及其地理上邻国之间的正向相关关系,远远大于我们所预期的、人均 GDP 的解释作用。这种结果是相当典型的,仅仅靠空间聚集的协变量,并不能完全去除研究中因变量的空间聚集效应。

假设在控制了国家的人均 GDP 以后,民主的分布依然表现出空间聚集效应,我们就需要通过可能的方法将空间依赖性纳入回归模型。与时间上的序列聚集情况一样,我们可以将空间自相关看作是外在干扰或本质现象。空间依赖性将使得对人均 GDP 的$\hat{\beta}$及其标准误的估计出现错误,这是由于在相互关联的个体之间,误差不能被视为相互独立。原则上,在用人均 GDP 对民主进行估计的例子中,这些问题可以通过某些考虑到空间误差相关性的估计值来解决;也就是说,残差的变化并不能完全被人均 GDP 解释。这种方法通常被称为空间误差模型,我们下面将会谈到这种方法。

然而,这里更大的关注点是什么影响了民主,而不仅仅是估计国家人均 GDP 与其民主发展之间的相关作用。如果一个国家的民主水平看上去与其邻国的民主水平相关,这就为我们提供了有关民主本身分布的信息,同时还让我们有机会了解空间依赖关系对民主的促进和阻碍作用。正因为如此,一个更可能和有趣的方法是将空间相关看作民主的本质特征,而不是一种统计干扰。

这里观测到的空间相关,表明观测值之间存在空间依赖,因此,国家 i 的民主值由于其周围国家 j 的民主程度不同而存在很大差异。与其将国家 i 的民主看作仅仅受人均 GDP 的影响,不如设计一个模型,将民主视为其自身人均 GDP

和周围国家民主水平的函数，定义为 $w_i . y_i$，这里对于所有同 i 相连接的国家 j，表示连接性的向量 w_i.（比如，矩阵 W 中的行 i）中的值都必须为非零值。前面我们曾提到，在表示连接的矩阵 W 中行经过了标准化，使得每一行的值加总为 1。

这种推导表明空间滞后因变量的模型具有如下形式：

$$y_i = \beta_0 + \beta_1 x_i + \rho w_i \cdot y_i + \in_i \qquad [2.1]$$

这里，空间滞后（ρ）的参数如果为正值，表明国家应该有更高的民主值，如果它们周围的国家的民主平均得分也很高。

这让我们容易把空间滞后 y 模型联想为类似的时间序列自回归模型，其中时间序列相关性的表示是通过在方程右边加入一个滞后的因变量 y_{t-1} 来估计其他协变量（比如 x_t）对 y_t 的影响。在空间滞后 y 模型中，系数 $\hat{\beta}_1$ 不同于 OLS 的回归系数，这是因为人均 GDP 对民主水平的估计作用大小控制了 y 的空间依赖关系，或者说国家 i 的民主水平的变化可以通过其他国家 j 的 y 值来解释。因此，在估计 x 变化带来的影响时，我们还需要考虑到空间上的相互影响。

表 2.1 和表 2.2 提供了考虑和不考虑 y 的空间滞后效应后 2002 年 158 个国家民主水平对人均 GDP 自然对数的 OLS 回归。在忽略空间滞后 y 的情况下，我们观察到：OLS 结果中人均 GDP 有非常大的正系数，为 1.68。相比而言，在空间滞后 y 模型中，人均 GDP 对数的系数为 0.76，比原来值的一半还要小，不过按照传统的显著检验标准，它还是远远大于 0。

表 2.1 没有空间滞后 OLS

OLS	$\hat{\beta}$	SE($\hat{\beta}$)	t Value
截距（Intercept）	−9.69	2.43	−3.99
人均 GDP 对数（ln GDP per capita）	1.68	0.31	5.36
$N = 158$			
Log likelihood ($df = 3$) $= -513.62$			
$F = 28.77$ ($df_1 = 1$, $df_2 = 156$)			

表 2.2 具有空间滞后的 OLS

	$\hat{\beta}$	SE($\hat{\beta}$)	t Value
截距(Intercept)	−4.98	2.07	−2.40
人均 GDP 对数(ln GDP per capita)	0.76	0.28	2.72
ρ	0.76	0.088	8.65
$N=158$			
Log likelihood ($df=4$) $=-482.48$			
$F=58.64$ ($df_1=2$, $df_2=155$)			

估计得到的空间滞后 y 不仅是较大正值(0.76)，而且按照标准它是高度统计显著的。这支持了我们的推测，即一个国家的民主水平和它的地理邻国的民主水平之间存在共变关系。实质上，该模型表示：如果一个国家的邻国的平均民主水平处于最小可能值(比如，−10)，与邻国平均民主水平为 0(这接近于 1945 年以来 POLITY 的历史平均得分)的国家相比，该国家的预期民主水平将少 7.6 分。相反，如果一个国家的邻国平均民主得分为 10，与一个邻国平均民主得分为 0 的国家相比，其预期的民主水平将高 7.6 分。这些估计结果反映了我们前面提到的民主的聚集效应。尽管大多数民主国家更可能具有高的人均 GDP，我们也观察到：2002 年，在一些民主聚集的地区，其人均 GDP 并不是很高，比如拉丁美洲，此外也有一些专制国家聚集区拥有较高的平均 GDP，例如波斯湾各国。

与表 2.1 中假设独立观测值的模型拟合度相比，表 2.2 中包括空间滞后 y 的模型对数据的拟合要好得多。该模型比假设观测值彼此独立的模型具有更高的 F 统计值和对数似然值。反过来，这让我们更加相信空间滞后 y 在识别民主分布中起到了重要作用，也就是说，不仅仅是国家人均 GDP 在起作用。然而，探索模型本身并不能令人完全相信空间方法的适用性。空间方法的优越性并不是因为仅仅起到探索性的作用，而是因为它为观测值之间的关联和相互反馈作用建立了一种看似合理的形式。

标准最小二乘法回归具有如下形式：

$$y_i = x_i\beta + \varepsilon_i$$

如果 ε_i 被分解为因变量的空间滞后项——它与因变量相关——和一个自变量的误差项，$\varepsilon_i = \rho w_{i\cdot} y_i + \in_i$，这就是空间滞后模型的形式：

$$y_i = x_i\beta_1 + \rho w_{i\cdot} y_i + \in_i$$

然而，如果换一种方法令 $\varepsilon_i = \lambda w_{i\cdot} \xi_i + \in_i$，得到：

$$y_i = x_i\beta_1 + \lambda w_{i\cdot} \xi_i + \in_i$$

这就是空间误差的形式。

下面我们将介绍空间滞后因变量模型；空间误差模型将留到第 3 章[9]讨论。

我们很容易将表 2.2 中的空间滞后 y 模型对人均 GDP 的系数估计，与表 2.1直接比较，并得出表 2.1 中人均 GDP 效应更大的结论。然而，这种解释并不正确。由于模型在方程 2.1 中加入了空间滞后之后变为一种自回归形式，因此 x 系数的作用反映了 x_i 对 y_i 的短期效应，而不是像不包含空间滞后 y 的 OLS 回归中 x 系数的净效应。由于 y_i 的值将影响其他国家 y_j 的民主水平，同时，这些 y_j 反过来影响 y_i，我们需要考虑到额外的效应，也就是 x_i 通过对其他国家民主水平的影响所导致的对 y_i 的短期影响。

这种解释类似于时间序列模型中协变量 x_t 的系数 β，此时方程的右边包括因变量的时间滞后变量 y_{t-1}，如下：

$$y_t = \beta x_t + \phi y_{t-1} + \in_t$$

在这个方程中，β 代表 x_t 对 y_t 的即时效应，但是反过来，它又会在下一个时间段上影响 y_{t-1}，同时 x_t 的长期效应也必须考虑净效应部分，该效应来自于自回归部分，或者说来自于滞后 y_{t-1} 的估计参数的影响。x_t 的长期效应为 $\beta/(1-\phi)$。当 ϕ 很大时①，长期效应 $\beta/(1-\phi)$将会显著大于 β。

沿用上面的类比，如果某一个国家 i 的人均 GDP 增加一个单位，这将对该国的民主水平产生直接影响 β_1。然而，方程 2.1 的模型表明：由于国家之间反馈

① 准确地说，应该是 ϕ 接近于 1 的时候。——译者注

作用产生的空间动态性，国家 i 的民主水平将对其邻国的民主水平产生影响。因此，i 的民主增长将影响邻国 j 的民主水平。同时，反过来，邻国的邻国也将受到影响，并且影响将扩大到所有相互连接的国家。一般来说，所有国家都会拥有一些邻国，因此最终所有的国家都会受到影响。但是注意方程 2.1 中包含了系统 y 中所有国家的民主水平，因此如果连接 i 的其他国家的民主水平提高，i 的民主水平也会提高。假设在一种实验状况下，某观测值受到一个外生的冲击，这种影响将通过观测值之间的相互作用引发一系列调整，并通过在系统中的循环产生回荡效应，直到生成新的稳定均衡（Cressie，1993；Lin，Wu，& Lee，2006）。

除了关注空间滞后 y 模型中 x_i 的估计系数，考虑均衡效应也很重要。可惜的是，空间滞后 y 的长期效应并不能像在时间滞后 y 情况中那样用简单的形式表示出来。我们会在后文讨论如何描述和估计空间滞后 y 模型中协变量的均衡效应。首先我们将讨论由于空间滞后 y 在方程右边所带来的内生性问题，以及与之相关的最小二乘法步骤中模型估计一致性的问题。

下面一节将利用矩阵代数讨论估计问题，以及为什么最大似然估计量（Maximum likelihood Estimator，MLE）适合用于估计空间滞后 y 模型。因为使用 MLE 本身并不需要了解本节中的所有内容，读者若对估计问题不感兴趣，可以跳过本节内容，直接进入下一节。

估计空间滞后 y 模型

在包含时间滞后项 y_{t-1} 的时间序列模型中，如果回归模型的残差不存在序列相关，时间滞后项 y_{t-1} 并不影响 OLS 的估计结果。更准确地说，假设模型被正确识别，包含滞后因变量的 OLS 模型并不导致估计问题。尽管在滞后因变量的优劣问题上已经存在大量争议，但这种争论取决于数据产生过程中的其他特定假设是否也合理。这里可参考基尔和凯利（Keele and Kelly，2006）的讨论。尽管 y_{t-1} 在时间上早于 t，y 在空间上的滞后却是同时存在的，并且来自于 y 自身。这种共时性（simultaneity）特征导致空间滞后 y 模型估计上的问题。为了

理解这一点，我们可以借助矩阵代数形式的空间滞后模型。根据安瑟林(Anselin, 1988)的表示法，空间滞后 y 模型可以写成：

$$Y = \rho \mathrm{W} y + \mathrm{X}\beta + \in$$

$$\in \sim N(0, \sigma^2 I)$$

这里 I 代表单位矩阵（$n \times n$ 矩阵中，对角线上为 1，其他地方为 0），$\in \sim N(0, \sigma^2 \mathrm{I})$表示误差服从正态分布，并且方差一致，即误差与协方差矩阵的积为 0。如果 $\rho = 0$，表明不存在空间依赖，将等式右边的第一项去掉后，我们就得到了标准的 OLS 回归模型。然而，如果 $\rho \neq 0$，由于共时性的存在，OLS 的估计将不会随着样本量的增加而收敛到它们的"真实"值。相反，在 OLS 形式中被忽略的反馈效应或依赖效应将增大，而不会随着样本量的增大而消失。实际上，这种效应明显依赖于样本量的大小和连接矩阵的具体形式。

如果靠 OLS 来估计空间滞后 y 模型存在问题，那么有没有其他估计方法呢？空间滞后 y 模型可以通过两步工具变量估计法来估计，比如，用外生变量 X、WX 和 $\mathrm{W}^2\mathrm{X}$ 作为空间滞后 y 的工具变量。在这里我们将不详细介绍如何用工具变量进行估计，而是关注如何在空间滞后 y 模型中进行最大似然估计。在模型被正确识别的情况下，这可以得到一致的和渐进有效的估计值。尽管由于方程右边 Wy 的存在，使得空间滞后 y 的 OLS 估计存在共时性问题，但是最大似然估计保持了渐进的性质，并且估计的不一致(inconsistency)或者偏误(bias)大小将随着实践中的具体情况发生变化。弗兰泽兹和海耶斯(Franzese and Hayes, 2007)利用蒙特卡罗模拟方法研究了不同估计值的性质，结果发现：在某些情况下，包含空间滞后 y 的模型的 OLS 估计值仍然比最大似然估计具有更小的均方误(Mean Squared Errors)。在小样本情况下，这将很难影响到空间表达式。

最大化空间滞后 y 模型的似然值是比较复杂的。为了说明这个问题，我们将空间滞后 y 模型写成如下形式：

$$\in = y - \rho \mathrm{W} y + \mathrm{X}\beta = (I - \rho \mathrm{W}) y - \mathrm{X}\beta$$

反过来，我们可以将估计量 β 写成：

$$\beta = (X'X)^{-1}X'(I - \rho W)y$$

当 ρ 未知的时候，要找出该模型的 β 估计值将是一件困难的事情，这是因为对数似然函数包含了 $|I - \rho W|$ 行列式。这个 ρ 的 n 阶多项式可以通过每一次迭代估算。然而，奥德（Ord，1975）表示如果 W 存在特征值（ω_1，…，ω_n），那么：

$$|\omega I - \rho W| = \prod_{i=1}^{n} (\omega - \omega_i)$$

反过来，这也表示：

$$|I - \rho W| = \prod_{i=1}^{n} (1 - \rho\omega_i)$$

奥德表明 W 的 ω_i 可以在模型剩余部分被估计之前就找到。

回顾方差一致的情况下，经典线性回归模型中的对数似然函数为：

$$\begin{aligned}\ln L(\beta, \sigma^2) = & -N/2\ln(2\pi) - N/2\ln(2\sigma^2) \\ & -(y - X\beta)'(y - X\beta)/2\sigma^2\end{aligned}$$

相比之下，空间滞后模型的对数似然函数为：

$$\begin{aligned}\ln L(\beta, \sigma^2, \rho) = & \ln|I - \rho W| - N/2\ln(2\pi) - N/2\ln(2\sigma^2) \\ & -(y - \rho W - X\beta)'(y - \rho W - X\beta)/2\sigma^2\end{aligned}$$

并假定 ω_i 在估计之前就已知，这样我们就可以通过最大化该函数，而很容易地得到空间滞后 y 模型的最大似然估计值。我们还需要确认系数不会导致爆炸性的反馈过程，因为这将导致协方差矩阵非正定。

尽管当前存在其他算法，但我们在使用空间滞后 y 模型的最大似然法时仍然最常采用奥德的方法，该方法可以消除复杂计算中的一个主要部分。然而，在采取最大似然法时，基本假设却发生了变化。在 OLS 中，误差要求服从正态分布，但是不一定需要数据。而在空间模型的最大似然估计中，数据被假设服从正态分布。

空间性间隔 y 模型的最大似然估计：以民主研究为例

在本节中，我们将列出民主的空间滞后 y 模型的最大似然估计，同时将其与相同模型的 OLS 估计结果进行比较。

首先执行这个命令的 R 编码为：

```
sldv.fit <- lagsarlm(democracy ~ log(gdp.2002/population),
  data = sldv, nb2listw(nblist), method = "eigen", quiet = FALSE)
summary(sldv.fit)
moran.test(resid(sldv.fit),nb2listw(nblist))
```

上述命令的结果见表 2.3。正如我们所见，对人均 GDP 的系数估计(接近 1.0)高于 OLS 估计结果(0.76)，而空间滞后 y 的参数 $\hat{\rho}$(0.56)却比空间滞后 y 模型的 OLS 估计更低。无论何种估计方法，我们的主要结论都是一致的，也就是：包括空间滞后 y 项将显著提高模型对国家间民主变化的解释能力。

表 2.3　空间滞后 y 模型的最大似然估计

	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截距(Intercept)	−6.20	2.08	−2.98
人均 GDP 对数(ln GDP per capita)	0.99	0.28	3.59
$\hat{\rho}$	0.56	0.08	7.43
$N = 158$			
Log likelihood ($df = 4$) $= -491.10$			

如果我们相信最大似然估计 MLE 相比于 OLS 更适用于估计空间滞后 y 模型，那么我们可以推断：OLS 估计低估了人均 GDP 的系数，且高估了空间滞后项的系数。但这种推测并不可以检验，因为我们并不知道“真实”的参数是什么，以及我们的模型和真实的情况有多接近，甚至不确定是否存在“真实”参数值。

针对残差自相关的拉格朗日乘子检验，更常用于检验空间模型中的残差。在本例中该检验得到的值为2.1，相应的概率为0.147，也就是明确拒绝了剩余残差间的一阶自相关的情况。除此之外，衡量残差的空间聚集程度的莫兰I估计值得到了同样的结果，鞍点修正法结果也是如此。莫兰估计得到的标准分为－0.46，这表示基于同样的连接矩阵W，我们拒绝了残差之间存在简单空间相关。如果我们对空间滞后 y 模型的OLS残差的空间模式进行检验，我们发现：大量证据表明残差之间仍然表现出很强的空间聚集效应，这时莫兰I为－0.17，相应的标准分为－3.21；鞍点估计结果也基本一致。负的莫兰I表明残差之间的排斥作用，这也支持我们的推测，即OLS过高估计了空间滞后 y 的作用，同时在对人均GDP效应的估计中过度矫正了空间依赖关系。我们强调应当谨慎使用残差的自相关检验，因为它们依赖于连接矩阵，而该矩阵本身在很多情况下都受到各种可能的识别方法的影响。我们在后文中还会讨论这一点。

空间滞后 y 模型的均衡效应

基于空间滞后 y 模型的最大似然估计，我们来考察人均GDP对民主影响的均衡效应。这需要我们考虑到自变量中一个国家 i 发生变化时，其他国家受到的影响。这将通过连接矩阵影响到其他国家的一系列变化，并最终通过空间滞后 y 项影响 y_i。

请记住：空间滞后回归模型可以被写成以下的矩阵形式：

$$y = X\beta + \rho W y + \in$$

将所有和因变量 y 有关的项移到左边，我们得到：

$$(I - \rho W) y = X\beta + \in$$

求解关于 y 的方程，再求期望值，我们可以发现，在这个均衡中，y 的期望值为：

$$E(y) = (I - \rho W)^{-1} X\beta$$

很明显，只有当 $\rho=0$ 的时候，$E(y)$ 才可能缩减为 $X\beta$。为了确定 y_i 的期望

值或者 x 的均衡效应，我们必须考虑到空间乘子 $(\mathrm{I}-\rho\mathrm{W})^{-1}$。这个乘子告诉：我们 x_i 的变化多少会"扩散(spill over)"到其他国家 j，并反过来通过 y 的空间滞后量影响到 y_i。这和列昂杰夫(Leontief，1986)在投入产出分析中应用的逆矩阵(inverse)方法相似，它用来估计一个部门的需求变化如何影响多个部门体系的总体产量。

因此，为了确定 x_i 中某个观测值的一个单位变化所带来的均衡效应，我们需要乘上向量 $\Delta x(i)$，而其他单位 j 的值都通过 $(\mathrm{I}-\rho\mathrm{W})^{-1}\beta$ 控制为常数。由于每个国家和其他国家之间的连接程度不同，并且高阶连接程度也不同，因此对一定 x_i 变化产生的影响将根据特定国家的改变而改变。假设我们有两个不相连的区域，彼此之间不存在桥梁连接，那么区域 1 的改变将影响到区域 1 中其他国家的改变，但是这些改变对区域 2 中的国家将没有影响。

描述均衡效应变化的一个有用方法，是考虑所有国家发生一定变化时产生的影响，并且考察每个具体国家估计值的分布。在这个例子中，我们得到的均衡效应均值为 1.09，它大约比表 2.3 中 $\hat{\beta}$ 系数估计值 0.99 所表示的人均 GDP 对数的短期效应高 10%。在个体上具体国家的均衡效应从低的 1.03(蒙古)到高的 1.24(巴布亚新几内亚)，后者比人均 GDP 的短期效应高 25%。很明显，我们不应该在还没有考虑到空间乘子以及空间个体之间的变动时就对空间滞后 y 模型中的协变量的作用进行推断。图 2.1 展示了估计效应的柱状图。

考察全向量 $(\mathrm{I}-\rho\mathrm{W})^{-1}\beta\Delta x(i)$，将有助于我们理解一个国家的人均 GDP 如何影响其他国家的民主预期值。比如以俄罗斯为例。表 2.4 列出了基于俄罗斯得到的 10 个 $(\mathrm{I}-\rho\mathrm{W})^{-1}\beta\Delta x(i)$ 的最高值，表 2.3 列出了该空间滞后 y 模型的估计值，连接矩阵用 W 表示。如我们所见，俄罗斯的潜在均衡作用为 1.09，这与模型中潜在均衡影响的中位数接近。其他国家的值表明：俄罗斯的改变将会对其亚欧邻国产生影响。为了弄清这些估计值在实际中表示的含义，我们可以重新回忆一下人均 GDP 对数的估计效应的系数。俄罗斯当前 GDP 每变化 10%(即 2279 美元)，将仅仅使其民主预测值的 POLITY 得分增加 0.1 分。在和俄罗斯发生相同的人均 GDP 变化情况下，即便是在该变化所产生的均衡效应最大的国家，其民主预测值也仅仅在此估计值基础上增加 0.02。这进一步强化

了我们的结论：即便某个国家的人均 GDP 发生很大差异，但根据我们的模型，这并不会使世界范围内的其他国家的民主预期水平产生很大的改变；同时，当空间滞后 y 模型考虑了相连国家之间民主水平的相互影响之后，人均 GDP 对数的影响将远远小于把各个观测值看作相互独立时的 OLS 结果。

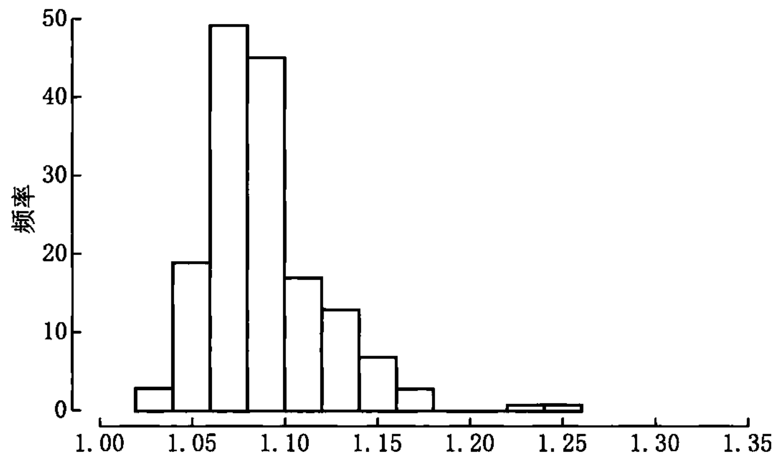


图 2.1 人均 GDP 自然对数均衡效应的直方图

表 2.4 俄罗斯等 10 个最高人均 GDP 得分国家的均衡作用

国　家	作　用	国　家	作　用
俄罗斯	1.09	爱沙尼亚	0.21
朝　鲜	0.24	挪　威	0.20
日　本	0.24	立陶宛	0.20
蒙　古	0.24	拉脱维亚	0.12
芬　兰	0.22	亚美尼亚	0.18

根据上文的空间滞后 y 变量估计值，我们利用下面的编码建构一个简化实验：

```
# Code to calculate equilibrium effect of changes in GDP per capita
# Create vector to store the estimate for each state
ee.est <- rep(NA, dim(sldv)[1])
# Assign the country name labels
```

```
names(ee.est) <- sldv$tla
# Create a null vector to use in loop
svec <- rep(0, dim(sldv)[1])
# Create a N x N identity matrix
eye <- matrix(0, nrow = dim(sldv)[1], ncol = dim(sldv)[1])
diag(eye) <- 1
#Loop over 1:n states and store effect of change in
# each state i in ee.est[i]
for(i in 1:length(ee.est)){
      cvec <- svec
      cvec[i] <- 1
      res <- solve(eye - 0.56315 * wmat) %*% cvec * 0.99877
      ee.est[i] <- res[i]
}
#Russia example of impact on other states (observation 120)
cvec <- rep(0, dim(sldv)[1])
cvec[120] <- 1
# Store estimates for impact of change in Russia in rus.est
eye <- matrix(0, nrow = dim(sldv)[1], ncol = dim(sldv)[1])
diag(eye) <- 1
rus.est <- solve(eye - 0.56315 * wmat) %*% cvec * 0.99877
# Find ten highest values of rus.est vector
rus.est <- round(rus.est, 3)
rus.est <- data.frame(sldv$tla, rus.est)
rus.est[rev(order(rus.est$rus.est)),][1:10,]
```

前面 OLS 模型的结果表明：人均 GDP 对民主预期水平的影响是相对有限

的。表 2.3 中空间滞后 y 模型的最大似然估计结果同样表明人均 GDP 的即刻影响相对较小。当我们考察人均 GDP 变化的长期均衡效应时，尽管该效应略微大但仍然有限。表 2.3 中的系数意味着一个国家的预期民主水平和其邻国之间存在什么样的关系呢？图 2.2 显示出模型中地理上预期的共变关系。在该图中，我们画出了因变量（民主，y）的期望值和其邻国之间民主水平（空间滞后变量 y^s 和自变量人均 GDP 的自然对数）的函数关系。该等高线图明确表明：人均 GDP 的作用很小，但是空间组成对国家的预期民主水平具有很大的作用。

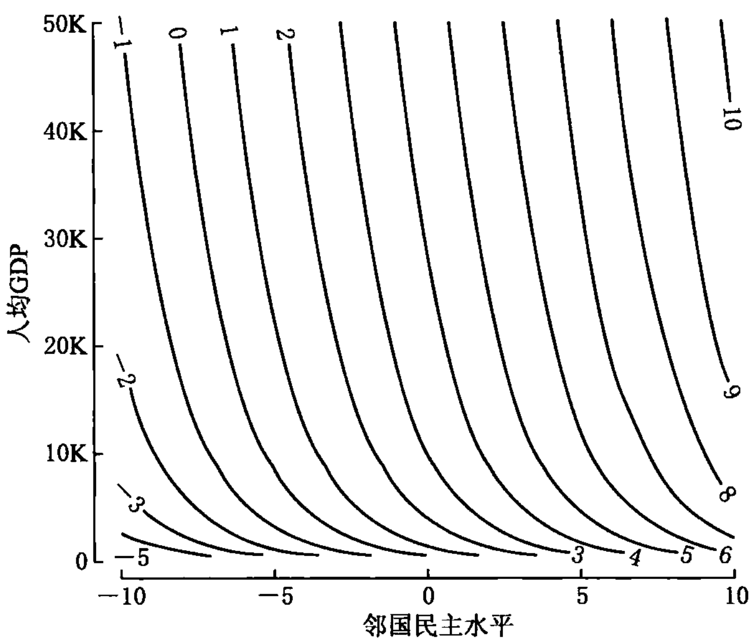


图 2.2 根据邻国民主水平和人均 GDP 对数预计得到的民主水平变化图

模型表明：控制了 GDP 水平后，一国的民主期望水平出现了很大的差异。如果一个国家的人均 GDP 取中位数，其邻国全部为专制国家（比如 $y_i^s = -10$），其期望民主得分约为 −4；然而如果它的所有邻国为民主国家（比如 $y_i^s = 10$），那么它的期望民主水平将接近 7。在这种情况下，尽管人均 GDP 在解释民主差异上的作用非常有限，但是国家的民主水平与其邻国之间却具有非常紧密的关系。

针对该结果的另外一种思路是，如果民主的改变来自于没有包含在模型系

统部分中的其他特征(比如某个国家 i 发生的震动)以及随之产生的对其他国家 j 的预测民主水平(在模型中表示为 $\hat{y}_j$)的短期影响,那么会出现什么结果呢?基于估计出的空间模型,这将对其他国家产生什么影响?很明显,这不仅依赖于 $\hat{\rho}$,而且还和建构空间滞后性的连接矩阵 W 的结构有关。

基于上面的空间滞后 y 估计,我们利用下面的编码来考察中国一个单位的 y 变化对系统中其他国家所产生的影响:

```
# Impact of change in $y$ to 10 in China
# China is observation 32
cvec <- rep(0, dim(sldv)[1])
cvec[32] <- 10
# Store estimates of change in China in chn.est
chn.est <- c(cbind (0, 0, wmat%*%cvec) %*%
     c(summary(sldv.fit)$Coef[, 1],summary(sldv.fit)$rho))
chn.est <- round(chn.est, 3)
# Find all states where non-zero impact
chn.est <- data.frame(sldv$tla, chn.est)
chn.est <- chn.est[rev(order(chn.est$chn.est)),]
chn.est[chn.est$chn.est>0,]
```

表 2.5 如果令中国的 POLITY 得分为 10,预测民主 $\hat{y}$ 的作用

国家(或地区)	作　用	国家(或地区)	作　用	国家(或地区)	作　用
中国台湾地区	1.88	老　挝	1.13	塔吉克斯坦	0.80
朝　鲜	1.88	吉尔吉斯斯坦	1.13	印　度	0.80
蒙　古	1.88	孟加拉国	1.13	越　南	0.80
尼泊尔	1.41	乌兹别克斯坦	0.94	阿富汗	0.80
不　丹	1.41	泰　国	0.94	哈萨克斯坦	0.70
巴基斯坦	1.13	缅　甸	0.94	俄罗斯	0.28

意大利投票率的空间依赖关系

锡恩(Shin, 2001)以及锡恩和阿格纽(Shin and Agnew, 2002, 2007a, 2007b)研究了意大利在过去的几十年中其政治活动的地理分布,并且发现在投票率和选举结果之间存在重要的空间动态关系。我们利用他们的数据来解释一种简单的观点,即投票率的空间变化可以通过意大利财富与收入的地理分布来解释。数据来自于意大利2001年国家选举和各省在1997年的人均GDP。这些数据包括选举中所有的477个学院(Collegi),或者叫单一席位选区(后面将用SMDs来表示)。

主要变量的地图

与空间回归分析中的第一步一样,这里我们先画出了选举投票率和人均GDP的地理分布关系(见图2.3和图2.4)。

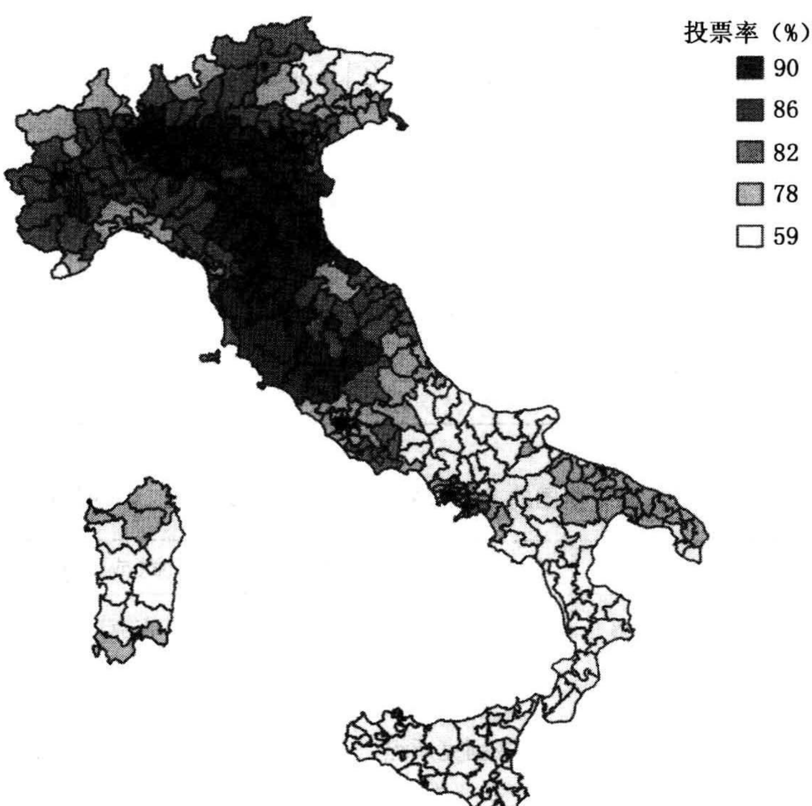


图 2.3 按学院(Collegio)划分的意大利选举投票率

投票率最高的是北部，尤其是北部最远的米兰(Milan)附近以及艾米利亚—罗马涅(Emilia-Romagna)和托斯卡纳(Tuscany)。罗马(Rome)和威尼斯(Venice)的投票率也很高。例如：在摩德纳(Modena)，投票率为90%。相反，在西西里(Sicily)投票率就停留在百分之十几；即便是在那不勒斯(Naples)市郊，选举投票率也仅仅约为60%。从人均GDP来看，意大利最富裕的地区是伦巴底(Lombardy)。北部最富裕选区的收入大约是南部最穷选区收入的1.5倍。这组探索性地图明显表现出了投票率和人均GDP的聚集效应。

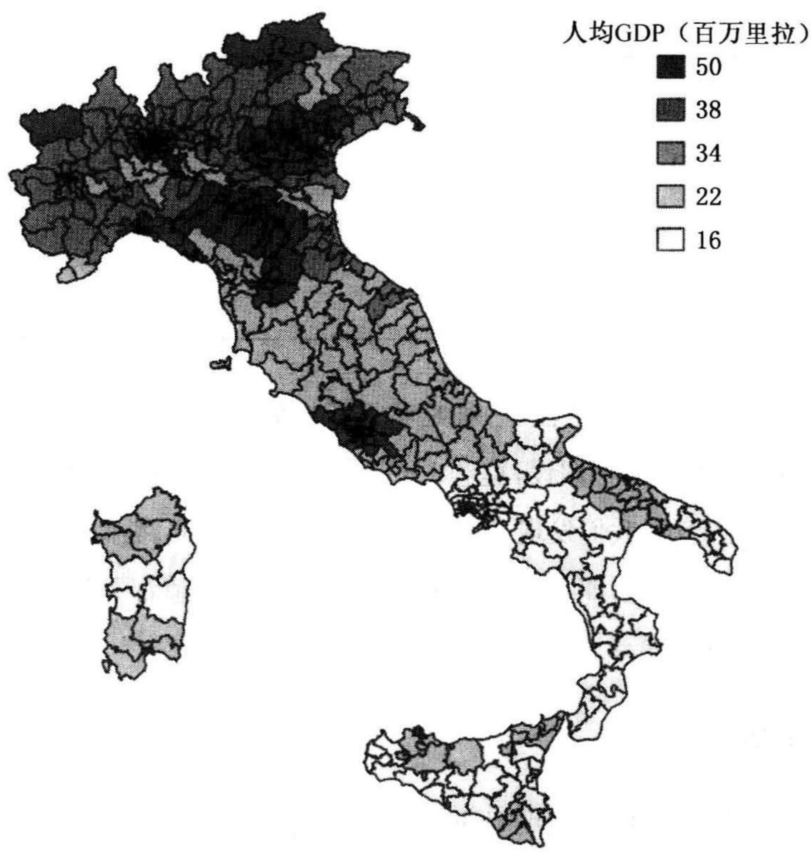


图2.4　意大利人均GDP，1997年数据

计算莫兰I统计量

在本节中，我们将利用莫兰I统计量这种更正规的方式来测算投票率和人均GDP的空间聚集效应。在测量意大利空间连接的第一步，我们先选取50千米

作为最近的邻近距离。通过计算各个区的图形中心，由此判断该区是否和其他区域的中心在 50 千米范围之内。总的看来，有两个米兰区域和 54 个其他区域相连：十区和六区。另外存在八个区仅仅和另外一个单一区相连，但这主要是边缘效应：比如特伦提诺阿托(Trentino-Alto Adige)，就位于阿尔卑斯山勃伦那(Brenner)山口和奥地利边境上。然而总体来说，各选区平均和其他 17 个相隔 50 千米以内的区相连。

对临近区域一览表的概括分析可以通过 R 很容易地得到。作为例子，我们用如下编码生成这些连接：

```
tr <- readShapePoly("turnout",
  IDvar = "FID_1", proj4string = CRS(" + proj = robin  + lon 0 = 0"))
dnn50km <- dnearneigh(coordinates(tr), 0, 50000)
summarize(dnn50km)
```

这里有两种莫兰计算方法，一种是基于随机假设，另外一种是基于正态假设。不管我们是基于何种假设，莫兰 I 统计量都表明数据中存在很强的空间关系。在这两种检验中，我们发现人均 GDP 的莫兰 I 为 0.86；类似的，投票率的相关上也具有很高的取值，即 0.79(两者都是)。所有这些值都是异常的，也就是表明人均 GDP 和投票率之间都存在很强的空间关联模式。

回归分析

投票率和人均 GDP 之间明显可能存在相关关系，但是投票率的空间聚集关系可以完全被 GDP 的地理差异解释吗？下面考察了投票率和人均 GDP 的简单函数模型。首先我们考查标准的最小二乘估计，结果在表 2.6 中列出。标准结果表明：在意大利，收入是投票行为的一个很强预测指标，对数人均 GDP 一个单位的变化(每百万里拉)将导致 14%的投票率变化。然而，我们得到回归残差空间模式的莫兰值为 0.47，这表明还存在没有被协变量解释掉的空间模式。

表 2.6 意大利投票率对人均 GDP 对数的 OLS 回归(1997)

	$\hat{\beta}$	SE($\hat{\beta}$)	t Value
截距	35.30	2.21	15.96
人均 GDP 对数	13.46	0.65	20.84
$N = 477$			
Log likelihood ($df = 3$) $= -1387.57$			
$F = 434.4(df_1 = 1, df_2 2 = 475)$			

然后我们用下面的代码,考察空间滞后 y 回归模型:

```
shin <- read.csv("italyturnout.csv",sep = ",",header = T)
sldv.fit <- lagsarlm(turnout ~ log(gdpcap), data = shin,
        nb2listw(dnn50km), method = "eigen", quiet = FALSE)

summary(sldv.fit)
```

表 2.7 列出了计算结果。人均 GDP 对投票率的作用没有上面的 OLS 结果那么"强",但是更可信。它表明收入的作用减小,但仍然是一个很强的作用。然而,空间滞后变量的作用相当显著。

表 2.7 意大利投票率对人均 GDP 对数的空间滞后回归(1997)

	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截距	4.70	1.66	2.80
人均 GDP 对数	1.77	0.48	3.66
$\hat{\rho}$	0.87	0.02	36.7
$N = 477$			
Log likelihood ($df = 3$) $= -1193$			

均衡分析

根据上面的方法。可以很容易地算出 477 个选区中每一个的均衡值,也就

是模型的预期值。这里我们并没有将它们列出来，而是用一个简单的实验方式：假设意大利最穷区域雷焦卡拉布里亚-斯巴里(Reggio Calabria-Sbarre)的人均 GDP 翻倍，这样我们计算出这一“情景”下的预期值和模型中观测数据的预期值的差异。这个差异对大部分选区来说都是不存在的，但是对其附近的 15 个选区来说，这种单一选区人均 GDP 的变化将导致期望投票率发生 1%或者更大的改变。与预期一样，最大的改变是在相邻的选区之间。图 2.5 表示了意大利投票率发生变化后的分布。



图 2.5 由于南部意大利单一穷困区(雷焦卡拉布里亚—斯巴里)人均 GDP 翻倍而带来的预期投票率增长

下面的编码表示如何基于事先得到的空间滞后 y 值，构造针对雷焦卡拉布里亚-斯巴里(432 个观测值)的实验(简略形式)：

```
# Extract estimated rho
rho <- coef(sldv.fit)[3]

# Extract estimated beta
beta <- coef(sldv.fit)[1:2]

# Create a X matrix
X <- cbind(1,log(shin$gdpcap))

# Create an alternative X matrix, changing value for
# Reggio Calabria - Sbarre (obs 432)
Xs <- X
Xs[432]<- log(35)

# Create an identity matrix
I <- diag(length(shin$gdpcap))

# wmat is the weights matrix
wmat <- nb2mat(dnn50km,style = "W")

# Find equilibrium effect by looking at
# the difference in expected value for the
# two x matrices
Ey <- solve(I - rho * wmat) %*% (X %*% beta)
EyS <- solve(I - rho * wmat) %*% (Xs %*% beta)
dif <- Eys - Ey
```

将不同权重矩阵引入空间滞后因变量模型中

我们以 2004 年美国总统选举为例，说明空间权重矩阵的作用[10]。从该数

据中，可以很容易提取出 XML 表，并且转变成 csv 文档。为了简化研究问题，我们将不考虑阿拉斯加和夏威夷，因为它们和其他所有的州都隔得足够远，所以并不会影响我们对区域数据的分析。我们感兴趣的主要变量，是小布什和克里的总选票在 48 个相连州与哥伦比亚地区中的份额。考虑到这是作为练习的一个例子，我们忽略每个州的提名投票数。然后构造一个小布什对克里的投票比，将此作为因变量。

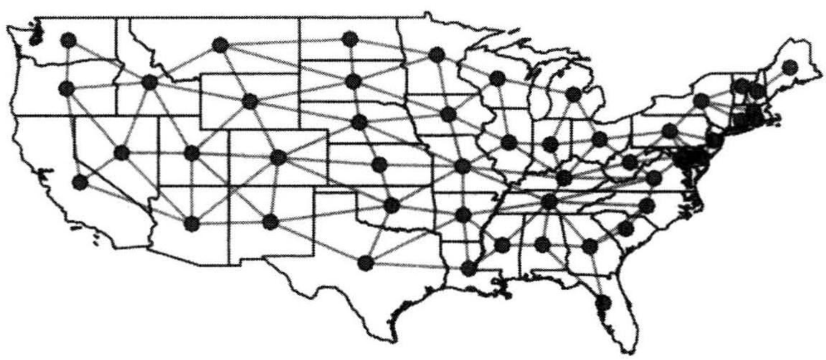


图 2.6　美国 48 个州按照接壤计算出的一阶连接图

为了找出空间类型数据中的自相关程度，我们构造出这 49 个政治和地理单位之间空间连接性的几种测量。第一种测量是相连州之间的简单测量。在这种情况下，华盛顿州与爱达荷州和俄勒冈州相邻，因为它们接壤。科罗拉多州与新墨西哥州、亚利桑那州、犹他州、怀俄明州、内布拉斯加州、堪萨斯州以及俄克拉荷马州接壤。在分布的另一端，缅因州只与一个州接壤。图 2.6 描述了这些情况。

绘制这个地图的编码样本如下：

```
library(maptools); library(network)
library(spdep); library(sp); library(rgdal)
setwd("...")

# read in 2004 presidential votes
presvote <- read.table("2004presvote.csv", sep = ",", head = T)

# read in shape files for 48 US States plus District of Columbia
```

```
# will create a MAP OBJECT
# use equal area projection(Robinson)
usa.shp <- read.shape("48_states.shp")

usaall <- merge(usa.shp$att.data, presvote,
      by.x = "STATE_NAME", by.y = "State",
      sort = F)

# Create a distance matrix from original polygon shape file
tr <- readShapePoly("48_states, shp",
  IDvar = "ObjectID", proj4string = CRS(" +proj +robin  +lon 0 = 0"))
centroids <- coordinates(tr)

# Create polygons in a spatial object
us48polys <- Map2poly(usa.shp,
      region.id = as.character(usa.shp$att.data$STATE_NAME))

# Create neighbors, list, and matrix objects from polygon
  centroids
us48.nb <- poly2nb(us48polys,
      row.names = as.character(usa.shp$att.data$STATE_NAME))
us48.listw <- nb2listw(us48.nb, style = "B")
us48.mat <- (nb2mat(us48.nb, style = "B"))

# plot the network among the centroids
colnames(us48.mat) <- rownames(us48.mat) <- usa.shp$att.dat
$STATE_ABBR
usa <- network(us48.mat, directed = F)

set.seed(123)
```

```
# plot network first; then add state boundaries
plot.network(usa1, displayisolates = T, displaylabels = F,
       boxed, labels = F, coord = centroids, label.col = "gray20",
       usearrows = F, edge.col = rep("gray60", 190),
       vertex.col = "gray30", edge.lty = 1)
plot(us48polys, bty = "n", border = "slategray3", forcefill = TRUE,
     xaxt = "n", yaxt = "n", lwd = .000000000125, las = 1,
     ylab = "", xlab = "", add = T)
```

下面我们用地图画出 2004 年总统选举中，小布什和克里在每个州的得票比。如图 2.7 所示，2004 年总统大选中各州之间的投票行为表现出很强的地理关系。

```
library(RColorBrewer)
# now plot the Bush: Kerry vote ratio
bk <- usaall $ Bush/usaall $ Kerry
# set up five categories and assign colors
breaks <- round(quantile(bk), seq(0, 1, 1/5), na.rm = TRUE), 1)
cols <- brewer.pal(length(breaks), "Greys")

# use findInterval to color states by bk variable
plot(us48polys, bty = "n", border = "slategray3", forcefill = TRUE,
     xaxt = "n", yaxt = "n", lwd = .000000000125, las = 1, ylab = "", xlab = "")
plot(us48polys, bty = "n", col = cols[findInterval(bk, breaks,
all.inside = T)], forcefill = T, add = T)

legend(x = c(-125, -115), y = c(27, 32), legend = leglabs
(breaks),
        fill = cols, bty = "n")
```

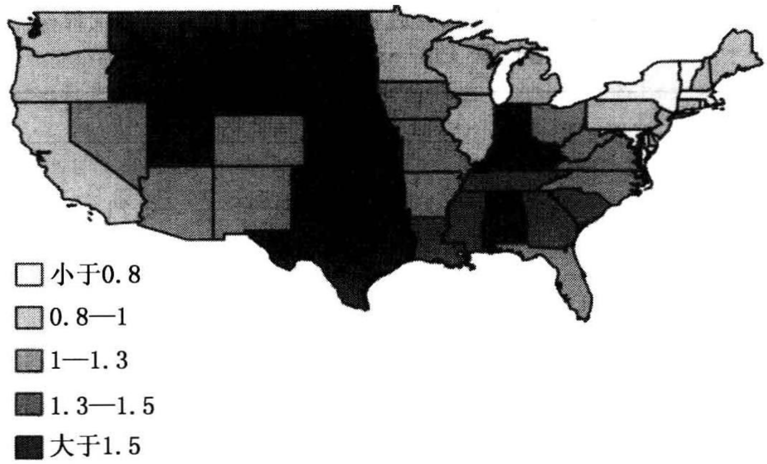


图 2.7　2004 年总统选举各州支持小布什对克里的投票比

莫兰$\mathcal{I}$同样在数字上表明了这种空间关系(见表 2.8)。

表 2.8　2004 年总统选举小布什—克里总投票数的自相关

数　　量	标准分	加权模式
莫兰(Moran's) $\mathcal{I}$		
0.39	4.7	边境接壤的州
0.49	5.7	最近的 4 个邻州
0.30	7.0	最近的 12 个邻州
吉尔里(Geary's)$\mathcal{C}$		
0.65	−2.7	边境接壤的州
0.65	−3.6	最近的 4 个邻州
0.69	−5.1	最近的 12 个邻州

各州的平均生产总值(类似于 GDP)来自于经济分析局的测量,该局是美国商务部的一部分。最新的可用数据来自于 http://www.bea.gov/bea/newsrelarchive/2006/gsp1006.xls,这也包括 1997 年到 2004 年各州生产总值的增长率。这些数据描绘了在 2004 年选举之前的 7 年中,各州经济的变化。我们用该数据作为协变量来解释 2004 年总统选举的投票情况。

我们建立了两个基本的空间连接矩阵,一个是根据是否接壤,另外一个是选择最近的四个邻州。对于每一个接壤编码,我们都估计了一个空间滞后变量的回归模型。结果见表 2.9。

表 2.9 2004 年美国总统竞选中各州小布什对克里的投票率对 GDP 增长率(1997—2004)的空间滞后回归

No.	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
是否与皇后地区(Queen)接壤			
截距	0.86	0.21	4.00
GDP 增长率	−0.05	0.06	0.85
$\hat{\rho}$	0.09	0.02	20.4
$N = 49$			
Log likelihood ($df = 3$) $= -25.63$			
4 个最接近的邻近州			
截距	0.63	0.23	2.72
GDP 增长率	−0.06	0.05	1.04
$\hat{\rho}$	0.60	0.12	18.4
$N = 49$			
Log likelihood ($df = 3$) $= -25.19$			

经验研究结果表明,GDP 增长率更高的州,在投票上更少支持小布什而更多支持克里。在接壤编码上,我们发现了很弱但正向的空间关系(0.09),但在利用四个邻州作为空间加权编码的时候,小布什—克里投票比却表现出更强的正向相关关系(0.60)。这两种不同的估计方式不仅在标准回归结果表上出现不同的结果,而且更重要的是它们将导致均衡值出现很大差异。如图 2.8 所示,对于这两种不同的加权模式,尽管其均衡效应正向相关,但它们的分布却完全不同。

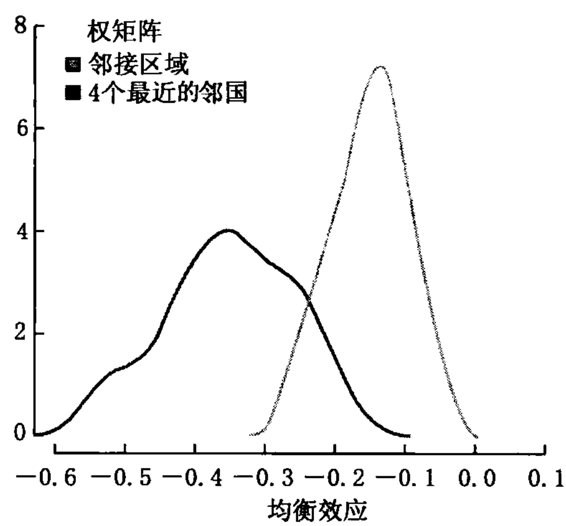


图 2.8 不同加权模式下空间滞后因变量的均衡效应的后验分布

边境接壤方法得到的均衡效应的均值(−0.15)比用四个邻近州作为估计得到的均值(−0.35)更小。该例子说明的首要问题是:加权矩阵在空间分析中起到了重要的作用,即便加权矩阵中很小的波动,也会对经验结果产生显著的影响。

空间滞后因变量与OLS中虚拟变量的比较

社会科学家们经常意识到:世界各个地区之间存在巨大的异质性,因此回归模型中每个国家各自的协变量并不能很好地解释空间差异。解决空间异质性的通常做法,是将各个不同地区作为虚拟变量加入模型。这种虚拟变量可以完全拟合不同地理区域的截距,从而考虑各个分散区域中因变量 y 的固定均值差异。这是在当前应用研究中最常用的处理区域异质性的方法,社会科学中有很多模型都是这样将"区域"分类作为虚拟变量加入模型的。同时,这样的模型也变得越来越普遍,因为分析者们越来越多地注意到合并数据的OLS(pooled OLS)估计可能没有考虑到各区域的重要差异。

比如,李(Lee, 2005)在一项民主和公共部门规模对收入不平等的影响的研究中,将非洲、亚洲和拉丁美洲作为区域虚拟变量拟合模型,并提出后两个区域相比于参照类(经济合作和发展组织)明显不同,并且这种不同不能通过模型右边国家方面的具体变量来解释。在民主研究中,布赫特(Burkhart)和刘易斯-贝克(Lewis-Beck, 1994)将世界体系中不同的民主水平作为虚拟变量,以此处理模型中的异质性问题,从而将国家区分为世界经济的中心(Core)、边陲(Periphery)和半边陲(Semi-periphery)。

在模型中加入区域虚拟变量的方法在社会科学中非常受欢迎,这也成为空间滞后 y 模型的一种替代方法。下面我们将区域虚拟变量加入模型中,作为原始OLS模型的替代方法,然后讨论这种模型和空间滞后 y 模型的关系。表2.10列出了将拉丁美洲和加勒比海、欧洲、撒哈拉以南非洲地区、中东和北非、亚洲以及大洋洲作为虚拟变量的模型。该模型忽略的区域,也就是参照类,为北美洲(即美国和加拿大)。不同区域的系数估计表明:在控制了人均GDP以后,某一地区的一个国家预测的民主水平相比于北美的差异。拉丁美洲和加勒比海地区、欧洲、大洋洲看上去和北美的平均民主水平之间不存在显著差异,但是撒哈拉

以南非洲地区和亚洲，尤其是中东，相比于北美，其平均民主水平要低很多。我们同时注意到：这里的人均 GDP 对数的估计系数要比将各个国家作为独立个体得到的 OLS 模型（也就是 1.68）低很多。事实上，该模型的人均 GDP 系数和上面空间滞后 y 模型中得到的均衡效应（也就是 1.09）非常接近。这也表明合并数据的 OLS 忽略了空间异质性，同时通过虚拟变量控制区域差异，起到了解决区域异质性的作用。此外这也说明之前人均 GDP 的作用被高估了。

表 2.10 包含空间虚拟变量的模型估计

No.	$\hat{\beta}$	SE($\hat{\beta}$)	t Value
截距	−1.89	5.06	−0.37
人均 GDP 对数	1.15	0.34	3.39
拉丁美洲和加勒比海地区	0.09	3.84	0.02
欧洲	−0.41	3.74	−0.11
撒哈拉以南的非洲地区	−4.71	3.97	−1.19
中东和北非	−11.77	3.85	−3.05
亚洲	−5.97	3.92	−1.52
大洋洲	0.90	4.72	0.19
$N = 158$			
Log likelihood ($df = 8$) $= -477.52$			
$F = 18.65$ ($df_1 = 7$, $df_2 = 150$)			

空间虚拟变量的方法适合替代空间滞后 y 模型吗？回答这个问题的一个可能的办法是考虑这两个模型的简约性（Parsimony）。尽管包含虚拟变量的 OLS 具有更高的对数似然比，但这增加了 6 个新的参数，或者说比空间滞后 y 模型多出了 5 个参数。同时，区域虚拟变量模型本身并没有解释为什么会存在区域差异，它仅仅是基于观察到的区域差异拟合不同的截距。如果一个区域因为其人均 GDP 变化而导致其民主水平发生了改变，这也不会改变其他国家的预测值，因为区域差异被看作是固定的，并且国家之间也不会相互影响。相比之下，空间滞后 y 模型仅仅增加了一个参数，就可以从实质上解释与某一国家相连的国家的民主水平 y 对这个国家民主水平的作用。如果不考虑简约性，而仅仅是为了提高拟合度，当然可以用区域虚拟变量拟合空间滞后 y 模型。但在这种情况下，模型仍然表现出残差的空间聚集效应，也就是得到正向且统计显

著的估计值 $\hat{\rho}=0.25$。如果用空间虚拟变量拟合空间滞后 y 模型，那么分析者需要考虑的，是固定区域差异的相关假设问题，而不是空间滞后 y 表达式中所产生的内生性问题。在很多情况下，如果连接矩阵 W 和区域划分非常相似的话，我们将很难分别估计区域虚拟变量的参数和空间滞后 y 的参数，这与回归的共线性问题类似。进一步而言，区域虚拟变量的构造基于指定国家为分散无关联的或者特定名称的区域，而空间滞后 y 项的构造基于连接矩阵 W，每个国家都有特定的连接。与分散无关联的识别方法相比，基于国家之间连接的方法的优势在于它不会将地理上隔得很远的国家比如希腊和爱尔兰划为同一个群，也不会将跨越了几个共同的地理区域的国家，比如土耳其和俄罗斯，作为同一个区域处理。

即使我们相信互斥的区域也可以用来构造观测值之间的连接，但区域虚拟变量方法并不总是空间滞后 y 模型的合适替代，而且它需要更多过度限制的假设。为了说明这个问题，假设一个回归中有 k 个不同虚拟变量 D_k 且满足 $y=b_1D_1+\ldots+b_kD_k+e$。与连接列表或者连接矩阵不同的是，在连接表或者矩阵中，i 不作为其自身的邻国包含在内，而在这里，每个区域同时包括 i 和它所有的邻国。但是如果每个区域内的样本数很大，我们可能得到 $\mathrm{W}y\approx b_1D_1+\ldots+b_kD_k$。这表示虚拟变量回归可以重新写作 $y=b_1D_1+\ldots+b_kD_k\approx \mathrm{W}y+e$。这样，虚拟变量回归模型就变成空间滞后 y 模型的特例，这仅仅是假设 $\rho=1$ 而不是估计实际参数就是 1(Lin et al.，2006)。换句话讲，空间虚拟变量假设每个区域内的所有的观测值都是同质并且相互连接的，但是空间滞后 y 模型却允许我们估计不同的相似程度。此外，空间滞后 y 模型可以方便地处理各种连接形式，而虚拟变量方法则假设群体之间互不相连。也就是说，群体内的每个分析单位相互关联，而群体之间没有关联；另外，分析单位也不能同时属于不同的群体。

第 3 章 | 空间误差模型

在第 2 章中，我们考察了空间滞后因变量模型，在该模型中因变量的"邻近"值对它本身具有直接影响。尽管这可能是处理空间依赖问题最常见并且是最有用的方法，但是，在连续因变量线性模型中，这并不是表示空间依赖关系的唯一方法。在本章中，我们将考察另一个替代概念，也就是说空间依赖关系来自于误差，而不是来自模型的系统部分。这种模型通常被称之为空间误差模型。我们同时介绍空间回归模型的一种重要扩展方法，也就是将空间误差模型扩展到度量学(metric)的距离概念，而不是仅仅停留在地理距离上。

空间误差模型

尽管空间滞后变量模型将空间依赖看作是本质现象，也就是说 y_i 受到其他国家 $y_j(i \neq j)$ 的值的影响，但是，空间误差模型却将空间相关关系看作一种干扰。这类似于统计学方法中经常将时间序列相关视为需要消除的问题，而且仅仅是一种估计问题。这种方法一般关注模型系统部分中被估自变量的参数，而忽略了数据产生过程中，数据相关性本身所反映出的意义。空间误差模型假设模型的误差是空间上相关的，而不是认为 y_i 对 y_j 产生直接影响。这种空间相关性的建立可以通过很多种方法来识别。这里我们仅仅关注一种基于空间体系编码的简单方法，也就是空间权重的方法；其他重要的方法包括考察地理统计协方差结构，但是本书将不介绍这些方法。根据之前的定义，如果令 w_i 代表向量 W，也就是测量 i 和其他单位 $j \neq i$ 的接近程度，我们可以将空间误差模型写成如下的形式：

$$y_i = x_i\beta + \lambda w_{i\cdot}\xi_i + \epsilon_i$$

这里我们将总的误差分解为两部分——即：ϵ 为空间不相关项，它满足正常回归假设中误差项空间上不相关的条件，ξ 为包含空间因素的误差项。参数 λ 表示在连接向量 w_i 中，相邻观测值的空间误差项 ξ 的相关程度。另外，我们还可以将空间误差模型按照第 2 章中定义的项，写作矩阵的形式：

$$y = X\beta + \lambda W\xi + \epsilon$$

$$\epsilon \sim N(0, \sigma^2 I)$$

如果相连观测值 i 和 j 的误差之间没有空间相关存在，那么空间误差参数 λ 将为 0，同时模型将简化为标准线性回归模型，也就是个体观测值之间相互独立，这样我们就可以按照传统方法估计 OLS 模型。然而，如果空间误差参数 $\lambda \neq 0$，那么我们可以判断相连观测值的误差存在空间依赖。这种结果可能仅仅是出于巧合，或者反映了模型系统成分的识别错误，尤其当被省略变量存在空间聚集的时候。社会科学家们通常希望发现正向空间相关关系。这意味着相似值之间存在聚集；也就是观测值 i 的误差的大小会随着其邻近观测值 j 的误差变化而系统性变化，这样 i 的更小/更大的误差就会向 j 的更小/更大①的误差靠拢。这种残差聚集效应违背了误差项相互独立的假设。

误差项的空间相关将导致什么结果呢？如果我们按照误差项相互独立的假设来估计 OLS，又会导致什么结果呢？如果 $\lambda \neq 0$，即便忽略了空间相关关系，OLS 系数估计结果仍然是无偏的。然而，系数估计的标准误将会是错误的。在 OLS 对方差的估计中，假设观测值相互独立。如果这不正确的话，那么，OLS 对方差 $\hat{\sigma}$ 的估计将低估了实际的方差，这与时间序列相关情况下的误差估计类似。这种情况之所以发生，是因为对方差的估计忽视了相邻观测值误差项的相关。此外，估计系数也并不必然是我们想要得到的有效估计值，也就是不“接近”真实值。之后，我们将返回来讨论空间误差模型的估计，下面我们先介绍如何解释该模型以及它和空间滞后 y 模型的关系。

① 原书这里为更小，而译者认为 i 的误差与 j 的误差大小应该是对应的。——译者注

空间误差模型和空间滞后 y 模型表面上看起来很相似，因为它们都表明了观测值之间的空间依赖性。然而，实际上这两种模型识别方法的实质意义却是不同的。空间滞后 y 模型是一种联立模型(simultaneous model)，观测值之间相互进行反馈：y_i 的值影响到 y_j 的值，反过来影响到 y_k 的值，然后再影响到 y_i。正如在第 2 章中所见，一个观测值 i 的自变量的不同取值，将通过相连的观测值而传播，而净影响来自于这些不同取值通过空间滞后 y 项，对其他相连观测值产生的影响。与之相反，在空间误差模型中，模型识别中的依赖关系仅仅来自于误差项。空间滞后 y 项的缺失表明 i 的自变量的差异并不会影响到与 i 相连的其他观测值的结果。因此，在空间误差模型的识别中，观测值的相互联系仅仅是因为未测量到的因素，也就是因为某些未知因素在距离上相关。

空间误差模型的最大似然估计

在空间滞后 y 模型的例子中，方程右边代表空间滞后作用的系数 ρ 是明确表明我们研究兴趣所在的参数。在空间误差模型中，λ 系数表明残差的相关，而不是明确的研究兴趣的协变量。如果我们仅仅是对估计 x 的 $\hat{\beta}$ 感兴趣，而完全忽略 λ，OLS 估计值将会是空间误差模型的无偏和一致性估计，这与空间滞后 y 模型不同。然而，报告的标准误却是不正确的，同时估计系数也不一定是有效的。这种问题可以通过利用广义最小二乘估计来解决，这类似于在存在时间相关的情况下用广义最小二乘估计值，这样我们先估计序列相关，然后试图转换数据以清除序列相关，从而满足普通回归的假设。通常这可以通过对空间连接矩阵特征值的最大似然估计来解决。

空间滞后误差模型的对数似然方程为：

$$\ln L(\beta,\sigma,\lambda)=\ln|\mathrm{I}-\lambda \mathrm{W}|-N/2\ln(2\pi)-N/2\ln(\sigma^2)$$
$$-(y-\lambda \mathrm{W}y-\mathrm{X}\beta+\lambda \mathrm{WX}\beta)'(y-\lambda \mathrm{W}y-\mathrm{X}\beta+\lambda \mathrm{WX}\beta)/2\sigma^2$$

如同空间滞后 y 模型的对数似然值一样，我们遇到如何计算行列式 $|\mathrm{I}-\lambda \mathrm{W}|$ 的对数的难题，这是一个难以估计的 n 阶多项式。然而，我们可以在此

依据奥德(Ord, 1975)的结果将这个行列式写成连接矩阵 W 的特征值 ω_i 的乘积：$|I-\lambda W| = \prod_i (1-\lambda\omega_i)$。由于特征值 ω_i 可以在最优化之前决定，该步骤可以同其他参数的似然估计分开(Anselin, 1988; Bivand, 2002)。这种估计可以通过常用的软件选项来执行，包括 R 中的 spdep。

以民主和发展研究为例

为了说明一个空间误差模型应用的实际例子，首先我们重新讨论第 2 章中关于民主和财富的例子。我们利用和第 2 章相同的数据，在变量的所有构造上面都参照前面的细节。表 3.1 展示了民主和收入例子中的三组估计。第 3 列的估计结果表明考虑到空间相关误差后的模型估计结果，而第 1 和第 2 列重复了第 1 章和第 2 章中 OLS 与空间滞后 y 模型的估计结果。

估计空间误差的 R 编码非常简单明了：

```
# data and variables as employed in chapter 2.
sem.fit <- errorsarlm(democracy ~ log(gdp.2002/population),
      data = sldv, nb2listw(nblist), method = "eigen", quiet = FALSE)
summary(sem.fit)
logLik(sem.fit)
```

如表 3.1 所示，空间误差模型的人均 GDP 对数的估计系数远远大于在空间滞后 y 模型中的结果，尽管还是没有不包含空间关系的 OLS 模型的结果大。直觉告诉我们 OLS 模型可能因为没有考虑到民主和人均 GDP 在相邻国家间的空间聚集作用，而高估了人均 GDP 的直接作用。这样估计结果也更不精确。同时，我们也可以将空间滞后项看作当观测值相互独立的情况下，OLS 模型中被忽略的变量。相比之下，空间误差模型纠正了人均 GDP 和民主的正向空间相关，而且这种纠正减小了 GDP 影响的估计系数。然而，空间误差估计假设模型中观测值的空间依赖仅仅来自于误差，或者说是模型系统部分没有考虑到的因素。

表 3.1 民主和人均 GDP 对数

变量	OLS			SLDV			SEM		
	$\hat{\beta}$	SE($\hat{\beta}$)	z Value	$\hat{\beta}$	SE($\hat{\beta}$)	z Value	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截距	−9.69	2.43	−3.99	−6.20	2.08	−2.98	−7.49	3.07	−2.44
人均 GDP 对数	1.68	0.31	5.36	1.00	0.28	3.59	1.39	0.38	3.66
$\hat{\rho}$				0.56	0.08	7.43			
$\hat{\lambda}$							0.58	0.08	7.60
N		158			158			158	
自由度(df)		1			2			2	
对数似然值 (Log likelihood)		−513.62			−491.10			−491.53	

注:SEM 表示空间误差模型;SLDV 表示空间滞后因变量。

相反,在空间滞后 y 模型中,一国人均 GDP 增长带来的净效应,一部分会通过反馈效应实现,因为 i 的即时效应对其邻国 j 产生影响,然后通过空间滞后项又影响 i,这带来的民主分数的变化又会影响其他国家,并通过系统中的反馈直到产生某种均衡。因此,在空间滞后 y 模型中,人均 GDP 的估计系数看上去比空间相关误差模型更小,因为它反映的是即时效应,而不是模型中的长期净“均衡”效益。

空间滞后 y 和空间误差的比较

由于这里两个空间参数 ρ 和 λ 远远大于它们的标准误,因此我们可以放心地得出结论,认为数据中存在很大的空间依赖性,并且认定假设观测值间保持独立的标准 OLS 回归结果是具有误导性的。但我们会问:究竟哪个模型更好,空间滞后 y 模型还是空间相关误差模型呢?从统计上很难区分出空间滞后 y 模型和空间误差模型的好坏。这两个模型并不相互嵌套,因此也不可能将一个模型看作另一个的子集,也就是说,我们不能通过在模型中加入更多限制条件来进行假设检验。尽管可以通过正式的检验方法来比较非嵌套的模型[11],然而,通常这些结论都是不确定的,而且很难给出一个模型优于另一个的有力证据。在这个例子中,我们可以看到两个模型的对数似然值非常相似,空间滞后 y

模型的对数似然值仅略小于空间相关误差模型。由于两个模型的参数个数相同,我们也不能说一个模型比另外一个更简约,因此,我们的经验并不能告诉我们这两个模型哪个对数据拟合得更好。一种方法是通过交叉验证(cross-validation)或者是样本外的预测检验(out-of-sample prediction test),但是这些方法超出了本书讨论的范围。

更重要的是,究竟是空间滞后 y 模型还是空间误差模型更合适,实际上是先于理论的问题,这应当通过具体的研究问题来考虑。如果我们期望看到——或者感兴趣的是——反馈效应,那么空间滞后 y 模型应当是一个更合适的模型。在民主的例子中,合理的预期是一个国家的民主水平会受到其他国家民主程度的影响(参见例如 Gleditsch, 2002a; Gleditsch & Ward, 2007)。相反,如果认为模型中误差的空间相关来自于系统成分中其他被忽略的特征,而国家民主水平之间没有扩散效应,这种说法就显得更不可信。因此,在这个例子中,我们相信:空间滞后 y 模型比空间误差模型更恰当。

更一般的原因是:社会科学对空间误差模型更没有兴趣。在我们看来,只有当研究者们相信误差项可以存在某些空间模式,但他们却不愿意或者是无法对误差的来源提供假设的时候,空间误差模型才更适用。这样做的原因在于,社会科学中的大多数模型在识别个体观测值特征的时候都很难抓住所有的空间聚集作用。因此,空间滞后因变量的识别问题仍然有很多工作需要做。但是如果在某个领域大部分的重要机制都已经明确并且在模型的系统部分完全识别出来,而误差项中仍然存在相关的话,这时用空间误差模型来纠正残差干扰就非常有用。总的说来,由于社会科学模型通常很少关注数据之间的依赖关系,空间误差模型可以大大改进当前的模型。

估计成对贸易往来中的空间性误差

为了找出空间误差模型更适用的一个例子,我们考虑它在成对贸易往来研究中的一项应用。成对(Dyad)表示两个个体组成的一对,结果变量可以是某些个体特征或者个体间互动的测量,在我们的例子中表示两个国家 i 和 j 之间的

贸易量。在一些情况下,我们可能希望区分 i 和 j 互动的方向,比如 $i \rightarrow j$ 表示 i 对 j 的作用。相反,没有方向的互动可以用下标定义为 $i \leftrightarrow j$。一个包含 n 个个体的系统将产生 $n \times (n-1)$ 个有方向的组对,当我们不区分往来或者互动方向时,将有 $n \times (n-1)/2$ 个无方向的组对。除了我们例子中的贸易以外,成对观测值在国际关系中非常普遍,比如当我们对估计某些特征如何影响到一个特殊事件或者行为诸如两个国家 i 和 j 之间的冲突发生的可能性感兴趣的时候。

国际关系中成对分析的传统方法是将个体互动看作这一对或者这两个个体特征的函数,同时在考虑到相关解释因素之后将两个个体视为互相独立。而空间误差模型则有利于我们处理成对观测值之间可能存在的依赖关系[12]。

国际贸易的黄金标准模型一直以来都没有发生过大的变动:这类似于牛顿的万有引力模型。贸易被看作交易国家经济规模的函数,但是和国家之间的"距离"成反比。当前的经验研究表明很多因素都可能影响国家 i 和 j 之间的贸易程度。经验研究中最常用的贸易模型也被称之为贸易的万有引力模型,它假定两个国家之间的贸易量 $T_{i \rightarrow j}$ 和它们的经济(GDP_i 和 GDP_j)、人口(P_i 和 P_j)的地理距离($D_{i \leftrightarrow j}$)乘积成比例。该模型一般在取对数后表示成相加的形式:

$$\log(T_{i \rightarrow j}) =$$
$$\alpha + \beta_1 \ln(GDP_i) + \beta_2 \ln(GDP_j) + \beta_3 \ln(P_i) + \beta_4 \ln(P_j) + \beta_5 \ln(D_{i \leftrightarrow j}) + \epsilon$$

这里各种量的系数($\beta_1, \ldots, \beta_4$)应该为正值,而距离的系数(β_5)为负值。芬斯特拉,罗斯和马库森(Feenstra, Rose and Markusen, 2001)以及罗斯(Rose, 2004)在最近的研究中都提供了相应的例子。

万有引力模型的核心并没有涉及政治内容,但是很多社会科学家都感兴趣的是,政治因素如何影响了贸易往来。比如波林斯(Pollins, 1989a, 1989b)认为政治关系可能会对贸易量产生强烈的影响,因为一个国家不太可能和与它政治关系不好的国家有很高的贸易量,或许因为商人们担心贸易受到政治因素的破坏,或许因为政府对敌对的国家的贸易采取相应的限制。莫罗,西沃森和塔瓦雷斯(Morrow, Siverson, and Tabares, 1998)认为民主国家更可能和民主国家进行贸易往来,同时,与其他国家相比,它们之间的军事冲突会更少影响到贸

易。这些经验分析都表明这些特征会影响贸易往来。

在贸易研究中一个常常被忽略的问题是，成对的观测值之间可能并不是相互独立的。尽管有很多研究考虑到这个问题，一对有顺序的观测值在时间先后上可能并不是独立的（见 Beck & Katz，1996），但是大部分研究都假设不同对的观测值在同一个时间点上可以被认为相互独立。然而，在贸易往来研究中，我们有足够多的理由相信这种假设不成立。因为每个国家将和众多国家组成不同的组对，所以成对数据的结果会很复杂。首先，$T_{i \to j}$和$T_{i \to k}$的贸易流通并不能看作相互独立，因为它们的贸易输出方相同。第二，通常从国家 i 到 j 的贸易流通（$T_{i \to j}$）将和反方向的从 j 到 i 的贸易流通（$T_{j \to i}$）呈正向关系。这种数据中还常常发现高阶依赖关系[13]。经济学家们通常会将 $T_{i \to j}$ 和 $T_{j \to i}$ 的值取平均，然后用分解方法分析这个三角矩阵，但这种方法却使得观测值之间的依赖关系更强。此外，众所周知，大部分报告中的贸易数据是基于其他贸易流通数据的插补估计方法得到的（例如 Rozanski & Yeats，1994）。这种插补方法可能导致数字之间的序列相关。比如，世界银行报告的贸易数据就同班佛定律（Benford's Law）预计的首位数字分布（distribution of first digits）存在显著差异，班佛定律是一种常用于检验数据质量和识别数字是否捏造的方法[14]。

该例子就适于使用空间滞后误差模型，因为我们认为某些成队组的误差项之间相互关联，而不是观测到的在贸易流通上相互关联。净流通量将取决于成对的国家数量，但仅仅依靠这一点还没有考虑到由于成对依赖性所导致的误差变化。前面我们讨论了两个个体之间地理上的距离和连接状态。在这里，我们将依赖结构定义为成对国家间拥有一个共同的成员国，但这种依赖结构并不是传统意义上的"空间"。不过这并不阻碍我们将空间概念应用到非地理距离概念中。在这个例子中，我们的加权方案是，如果成对国家中包含 i 或 j 中的任何一个，则被看成是和 $i \to j$ 这对国家相连。有关"距离"替代概念的更多讨论，可以参见贝克，格里蒂奇和比尔兹利（Beck，Gleditsch and Beardsley，2006），迪沃斯和伊萨德（Deutsch and Isard，1961），以及洛夫达哈（Lofdahl，2002）的研究。

在经验应用举例中，我们参考格里蒂奇（Gleditsch，2002b）研究中使用过的欧洲和非洲成对贸易数据。具体来说，我们用 $T_{i \to j}$ 表示国家 i 到 j 的输出量。非

洲和欧洲的样本为我们提供了很有趣的比较，包括数据质量的变化，我们预计欧洲的贸易数据将比非洲的数据更精确，因为它们的基础设施以及经济活动监测能力都存在差异。该例子中所有的数据来源于 1998 年。在我们的样本中，贸易流动的“观察”数据来自于国际货币基金组织及其他国际机构，它们占欧洲所有成对数据的 75%左右（比如格里蒂奇 2002b 数据中原始编码以 0 或 2 开头的数据）。然而，对非洲而言，利用官方报告的数字我们将仅仅得到贸易流通数据中 15%的成对数据。在该例子中，对于欧洲我们将仅仅使用官方报告的数据，而在非洲贸易流通分析中，我们将使用所有可能的、甚至是有争议的数据来源作为估计。

标准的“万有引力模型”的变量包括经济规模，两个成员国的人口（数据来源于 Gleditsch, 2002b）和它们首都之间的距离。此外，我们的模型参考了现有文献中贸易的政治决定因素，并包括两个国家政治取向的相似性，这种相似性的测量是通过两个国家在联合国投票记录中的 S 相似性得分得到的（见 Gartzke, 1998; Signorino & Ritter, 1999）。民主的测量来自于 POLITY Ⅳ 数据。我们对数据进行适当的调整，包括对没有被纳入到自由屋（Freedom House）原始数据中的国家的 POLITY 数据也进行了估计[15]。根据贾格斯和格尔（Jaggers and Gurr, 1995）对制度化民主测量的 21 点量表，我们选取了两个数值中较低的那个，经过重新调节比例，使得所有的值都为正。最后，我们也考虑了成对的两个国家是否曾卷入军事化国际争端（见 Jones, Bremer and Singer, 1996）。

```
source("chapter3data.R")
tab3.sem <- errorsarlm(logtrade ~ logdem + logapop + logbpop +
        logargdppc + logbrgdppc + logs + logdist + logmid,
        data = logdat98, na.action = na.omit,
        nb2listw(dlist, style = "W"), method = "eigen")
summary(tab3.sem)
logLik(tab3.sem)
```

表 3.2 和表 3.3 分别列出了欧洲和非洲成对贸易往来的 OLS 和空间相关误差模型估计结果。从 $\hat{\lambda}$ 可以看出，非洲和欧洲样本中的成对贸易伙伴之间都存在很强的空间正向相关。同时，通过比较 OLS 和空间误差模型估计结果可以看出，当我们考虑到成对成员的残差空间相关，而不是把它们视为相互独立的观测值时，原有文献中强调的贸易变化的政治决定因素发生了很大的改变。尤其是当我们考虑到成对成员的空间相关关系之后，MID 所起的负向作用的估计系数在欧洲样本中降低了大约 25%，在非洲样本中降低了大约 40%。在非洲样本中，民主的估计系数降到了原有大小的 1/4，而在欧洲样本中该系数增大了约 15%。此外，空间误差模型中的个体系数估计值的标准误一般比 OLS 结果更大，这表明如果将个体成对观测值视为相互独立，由于模型中错误的标准误可能导致对估计结果过于肯定。更一般地说，尽管按照惯例标准，系数中并没有从“显著”变为“不显著”的情况发生，但很多独立成对个体假设下得到的明显结果，在考虑到空间依赖关系之后，变得不那么稳健。

表 3.2　出口，欧洲：$T_{i\to j}$

变　量	OLS			SEM		
	$\hat{\beta}$	SE($\hat{\beta}$)	z Value	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截距	−32.70	0.67	−48.82	−33.94	1.71	−19.90
民主对数	0.38	0.06	5.93	0.43	0.10	4.38
i 人口对数	0.86	0.02	40.37	0.89	0.03	31.46
j 人口对数	0.75	0.02	34.93	0.77	0.03	27.33
i 人均 GDP 对数	1.54	0.04	35.23	1.56	0.06	17.35
j 人均 GDP 对数	1.01	0.04	23.07	1.03	0.06	7.66
S 对数	0.33	0.05	6.92	0.35	0.05	7.69
$i\leftrightarrow j$ 距离对数	−0.34	0.01	−24.33	−0.34	0.01	−25.83
$i\leftrightarrow j$ 争端对数	−1.94	0.27	−7.14	−1.48	0.29	−5.01
$\hat{\lambda}$				0.98	0.01	73.73
N		1500			1500	
自由度(df)		8			9	
对数似然值 (Log likelihood)		−2324.8			−2239.668	

注：SEM 表示空间误差模型。

表 3.3 出口，非洲：$T_{i\rightarrow j}$

变 量	OLS			SEM		
	$\hat{\beta}$	SE($\hat{\beta}$)	z Value	$\hat{\beta}$	SE($\hat{\beta}$)	z Value
截距	−7.41	0.33	−22.38	−7.47	1.45	−5.16
民主对数	−0.04	0.04	−1.08	−0.01	0.05	−0.15
i 人口对数	0.26	0.01	20.51	0.26	0.02	14.45
j 人口对数	0.23	0.01	17.81	0.23	0.02	12.55
i 人均 GDP 对数	0.38	0.02	17.96	0.38	0.03	12.78
j 人均 GDP 对数	0.31	0.02	14.82	0.31	0.03	10.55
S 对数	3.41	0.40	8.50	3.43	0.47	7.24
$i\leftrightarrow j$ 距离对数	−0.17	0.01	−20.81	−0.17	0.01	−22.21
$i\leftrightarrow j$ 争端对数	−0.71	0.18	−3.85	−0.42	0.18	−2.37
$\hat{\lambda}$				0.99	0.01	124.2
N		2550			2550	
自由度(df)		8			9	
对数似然值 (Log likelihood)		−3096.2			−2945.9	

小结

在本章中，我们介绍了有关空间依赖关系的空间相关误差模型。在通常情况下，由于我们很难(尽可能地)完全基于统计标准判断空间相关误差模型和空间滞后 y 模型孰好孰坏，所以研究者们应当考虑这两种模型哪种能够为空间依赖关系提供最可信的解释。我们前面已经提到：空间滞后 y 模型更适合于处理当因变量的邻近值的变化对该个体因变量产生直接影响的情况，而空间相关误差模型更适用于当我们确信模型系统部分某些未观测到的特征，可能导致模型的误差出现空间相关模式的情况。在成对个体相互依赖的例子中，某个国家与很多观测值构成了不同的组对，这也表明：空间依赖概念可以从地理距离扩展到度量学的距离。

第 4 章 | 扩展

本书前面的章节讲解了将空间类型引入社会科学数据中进行分析的必要性和益处。我们介绍了如何将这种方法纳入常见的线性回归框架——空间滞后因变量模型和空间误差模型。在空间滞后因变量模型中，与 y_i 相连的单位将对 y_i 产生影响；在空间误差模型中，相连观测值的误差之间存在空间相关。这两种是最广泛使用的空间回归模型，它们有很多广泛应用。然而，空间回归模型也有很多其他的类型，以及在很多我们没有提及的情况下的扩展应用，并且我们前面的关注也仅局限在连续变量的横截面数据上。在本节中，我们列出了一些空间回归模型的扩展应用，以及空间分析可能面临的棘手问题。尽管我们的回顾非常简略并且不能提供涵盖这些扩展和替代方法的实际例子，但是我们提供了进一步阅读的参考建议。比万德，佩勃斯玛和戈麦斯-卢比奥(Bivand, Pebesma and Gomez-Rubio, forthcoming)提供了这些方法基于 R 统计软件的教学材料。

识别连接性

如何建构和处理观测值之间的连接，是分析中研究者所面临的一个关键问题。大多数空间回归模型的应用都事先假定了观测值之间的连接图。如何建立这些连接取决于有关观测值在实际中如何关联的理论或者直觉。在实际操作中，这些建立方法都是出于方便考虑或者是基于最新的常用方法。研究者们需要注意的是，选择不同的连接方法和编码方法可能意味着认识世界的不同观点。不同的结果可能造成个体直接连接形式的差异，但这并不是什么大问题。更微妙的地方在于这些选择也会影响空间结构中的空间乘子和模型中的协方

差结构(Wall, 2004)。即便是在地理距离的连接情况下,同样的空间布局可能由于研究者不同的决定而产生不同的连接结构。为了说明这一点,我们可以参考三种常见空间编码的差异:车(Rook,共同边界)、象(Bishop,共同顶点),或后(Queen,同时包括边界和顶点),见图 4.1 所示的美国局部地图。科罗拉多州和犹他州是具有共同边界和共同顶点的邻州。科罗拉多州和亚利桑那州没有共同的边界,但有共同的"顶点"。在当前的世界国家政治边界地图中,我们只发现了一个类似的情况:非洲西南的卡普利维地带(the Caprivi Strip)。

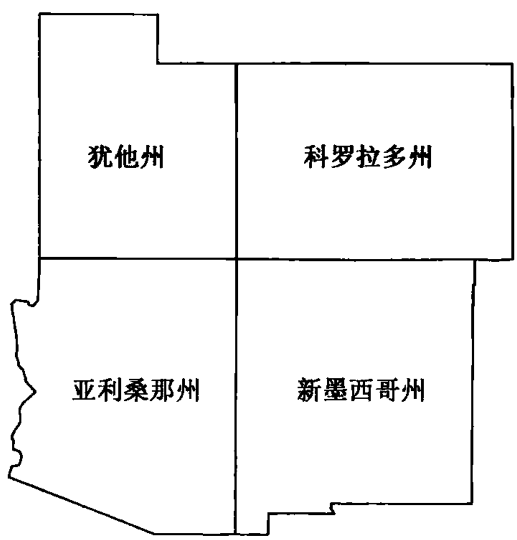


图 4.1 美国四角区域

更典型的情况是:如果个体距离在某个范围内,也就是行政中心或者地理质心或中心之间的最短距离,研究者们就将它们视为"相近"。格里蒂奇和沃德(Gleditsch and Ward, 2001)讨论了常用的中点测量存在的一些问题,也就是当个体的行政中心和边界相差很远或者是个体的奇怪形状导致中心不在版图之内的情况。划定过于窄的范围可能生成很多岛屿,这个问题我们在第 1 章新西兰的例子中提到过。从澳大利亚阿利斯斯普林斯(Alice Springs)到新西兰克赖斯特彻奇(Christchurch)大约为 4100 千米,相当于巴黎到达累斯萨拉姆(Dar es Salaam)的距离。这表明如果我们用澳大利亚和新西兰版图中心的距离作为标准并用到其他国家的话,那么大部分非洲国家、中东和亚洲国家都将成为法国

的邻国。划定标准过大的话将导致所有个体都相连。图 4.2 表明两幅国家连接图中，当中心连接距离从 400 千米(图 a)变为 4000 千米(图 b)时，连接密度的急剧增加。如果针对特定的个体设定连接规则，或者选取 k 个最相邻的观测值相

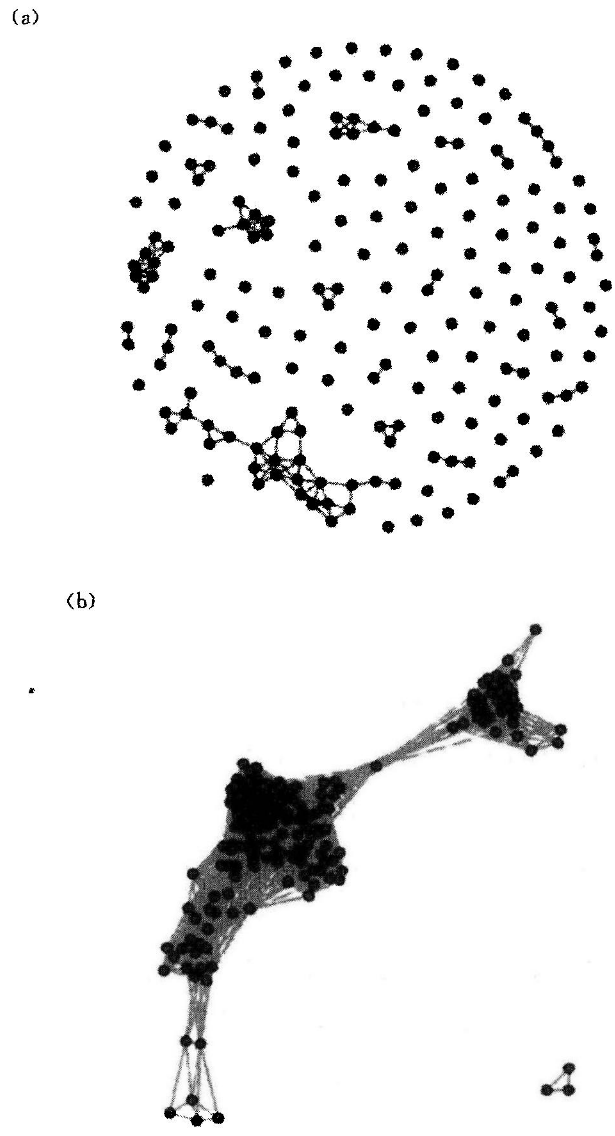


图 4.2　国家之间的连接，4000 千米和 400 千米的距离，基于版图中心间的距离：(a)400 千米；(b)4000 千米

互连接，则会引出另外一个问题，即为什么不对其他的个体也使用同样的标准。然而，这种特定规则在应用研究中不仅可能有用，而且是必要的。总的说来，对连接编码的不同选择总会对经验结果产生实质性的影响，因为通过不同的网络扩散将带来不同的结论。

基于非地理性测量的连接，比如贸易往来，可能存在其他的问题。更确切地说，如果非地理性的距离测量是基于实际空间回归模型中已有的变量，那连接将可能不是外生性的，这将导致模型识别和估计上的问题。研究者需要设计出与空间互动过程相匹配的连接矩阵。尽管拟合优度和交叉验证方法有助于剔除错误选择，但是连接性的设计其实是一个理论问题，并不可能使用简单的诊断或者探索方法来定义唯一"正确"的连接方式。我们还要强调的是，连接性识别上的困难也增大了检验空间依赖关系的零假设的难度，因为拒绝零假设只是针对某种特定连接方式而言的。

处理连接性

在识别出连接矩阵以后，如何在分析中处理连接矩阵则是另外一个问题。我们应当对所有的连接赋予相同的权重，还是应当根据观测值大小或重要性给予一些观测值不同的权重？在本书的例子中，我们假设俄罗斯和爱沙尼亚和与它们相连的国家之间的权重相同。然而，并没有谁规定所有的连接就应当使用相同权重。针对具体研究问题，研究者可能尝试不同的加权方法。

在回归模型中，我们仅仅考虑了行标准化矩阵 W 的情况，也就是所有的连接权重相加为 1。这种标准化的优点在于空间滞后 y^s 和 y 具有同样的潜在测量标准或单位。然而，标准化是否合理还是应当具体问题具体分析。举例来讲，默多克等人(Murdoch et al.，1997)关注过一个国家的污染排放量如何受到其他国家排放的影响。这个问题涉及总的污染排放量；这时利用相连国家的数量来对连接矩阵进行标准化可能就不适用了。

分析者可以将空间统计文献中惯用的做法作为参考，并仔细分析它们在自己的研究中是否行得通。比较有用的方法是多考虑几种替代方法。

一个对多个的连接

到目前为止，我们讨论了具有单一空间依赖项并表示成单一连接矩阵的例子。在很多情况下，可能会出现多种连接网络或者依赖形式的情况。通常可行的方法是根据地理距离或者其他政治网络（比如贸易合作、文化相似性，或者种族划分、职业划分）考虑几种不同的连接方法（见 Beck et al.，2006；Lacombe，2004；Lin et al.，2006）。直接影响不仅可能来源于一阶连接，也可能是高阶连接。图 4.3 画出了前面例子中提到的 158 个国家的一阶和高阶连接。

空间滞后 y 模型可以被推广到包括两个（或更多）不同连接矩阵 W^A 和 W^B 的形式，并通过如下表达式分别估计参数 ρ_1 和 ρ_1 各自的相对影响。

$$y_i = x_i\beta + \rho_1 w_i^A y + \rho_2 w_i^B y + \epsilon$$

扩展后的空间滞后 y 模型变得比标准空间自回归模型更难估计。假如这两个矩阵差别足够大并且不包含重复信息，那么这个模型便可以估计。如果这两个矩阵太相似，那么就会出现诸如经典回归模型中的共线性问题。前面讨论过的最大似然法也可以应用到这种情况（尽管还没有在 R 中实现）。这种模型也可以用工具性变量来估计。

推论与模型评估

和社会科学中的大多数数据一样，空间数据也不是来自于随机样本。空间分析需要一个相对完整的空间覆盖，这是因为包含太多缺失值的数据可能使对空间聚集影响或者相邻单位之间影响的统计推论变得毫无意义。虽然经典统计推论在很大程度上是基于渐进性假设，但在很多空间情况中这一点却很难证明。实质上，这要求某地区邻地的数量不会随着该地区的大小而剧烈变化。即便满足这一点，研究中的空间数据看上去并不像一个样本，而是像截面上的某个地区或者是整个世界。这使得我们需要找出符合研究背后社会过程的合理的模型，也就是原则上产生观测数据的模型。经典的方法可能是基于一种概括

(a)

(b)

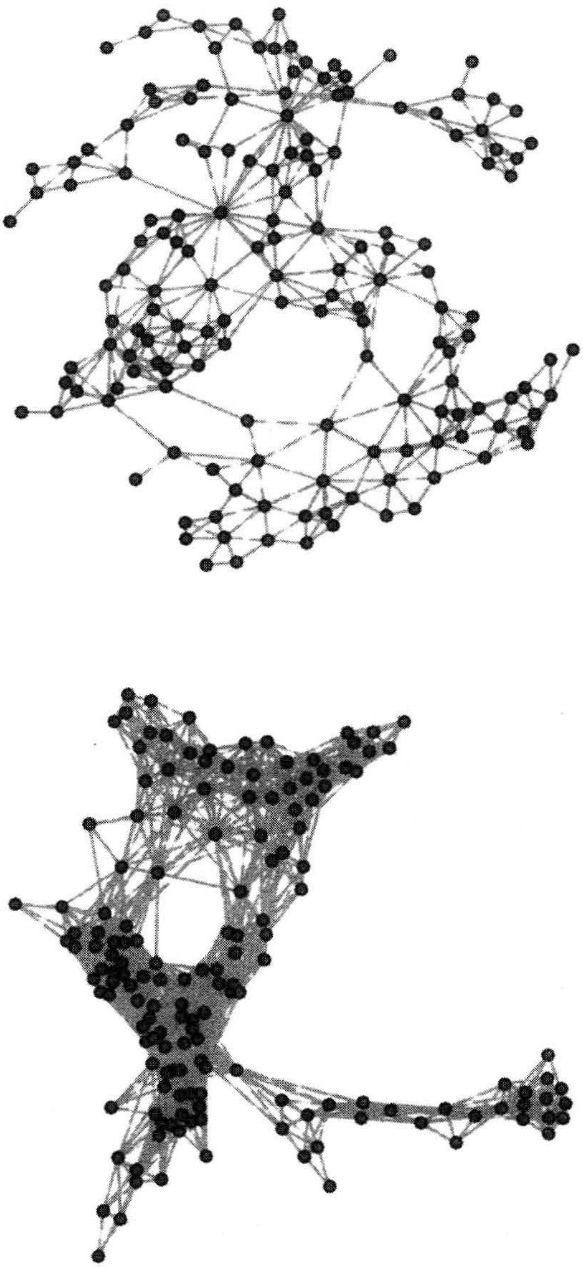


图 4.3　158 个国家一阶和二阶连接，基于最近邻国距离为 200 千米：(a)一阶；(b)一阶和二阶

性的“超总体(super-population)”概念，也就是观测到的空间类型在现实中的体现，但是这种概念在贝克、韦斯特和韦斯(Berk, Western, Weiss, 1995)那种所

谓的“明显的总体(apparent population)”的空间分析中并不适合。

这种难题的一个可能的解决办法是以探索性的态度看待估计结果，并且按照格斯尔(Geisser, 1974, 1975)的方法，用空间回归估计中没有用到的数据进行交叉检验，来检查模型的估计好坏。在空间背景下，这种方法可以通过利用后一个时间段或不同空间领域的观测数据来实现。比如，比万德(Bivand, 2002)就将数据分成两个地理区域，根据它们对数据另一半的观测值的预测能力，来评估不同建模方法的好坏。

离散和潜变量

我们前面假设 y 应当是连续变量。然而社会科学家感兴趣的很多现象都是离散的，也就是说以二分的或者计数的形式出现。另外，看到的结果也可能来自于潜在社会过程的一部分。这就好比线性回归只是这些数据的次优选择一样，这里估计的模型也不一定最适合这些数据。然而，将滞后因变量或者自回归过程扩展到二分情况或者计数数据也是可能的，比如在自相关逻辑斯蒂(autologistic)模型中，相邻个体的 y 值会影响 $\Pr(y_i = 1)$ 的情况(Besag, 1972, 1974; Christensen & Waagepetersen, 2002; Huffer & Wu, 1998; Ward & Gleditsch, 2002)。估计这些模型将比估计连续变量的情况复杂得多，因为 y 同时出现在方程的两边，这使得似然值很难处理。传统的方法在估计中将相连观测值的 y 视为固定，但是当前的计算能力使得利用模拟方法估计整个似然值成为可能。

空间异质性

通常回归中的主要效应为固定效应(fixed effects)，这表明自变量和因变量之间的关系是无处不在的。然而这种关系可能在世界上某个地方不同于其他地方，空间异质性指的就是这种效应和地理条件有关。空间异质性既为我们提供了了解研究现象的机会，也给研究带来了困难。一方面，它让我们可以分解回归结果，使得这些结果能更好地反映不同地区的情况。另一方面，它又违背了标准回归中所有分析数据方差相同的假设。地理加权回归，即 GWR(Geo-

graphically Weighted Regression)，作为一种数据探索的开创性方法，让我们利用某一位置的相邻观测值作为权重，估计每一个地理位置的回归系数。布伦森、弗泽林哈姆和查尔顿(Brundson, Fotheringham and Charlton, 1996)发展了空间分析中的这种方法，更多详细内容可以见弗泽林哈姆、查尔顿和布伦森(Fotheringham, Charlton and Brundson, 2002)的研究。在政治科学中，一个最近的例子来自卡尔沃和埃斯科拉(Calvo and Escolar, 2003)；人口学中一项有意思的应用是厄舍克和皮纳斯古鲁(Işik and Pinarcioğlu, 2007)的研究。

点和地理统计数据

前面的方法都是将地理空间看作可以划分的。比如国家被看作一个个格子，表明每个国家都可以在地图上找到一个方格，没有哪个国家占有超过一个方格的位置。在很多数据中，这种方法都是有用的，但并不是所有的现象都可以被视为区域或者格子，通常数据也不是按照这种格式来表现的。事实上，很多类型的数据都是地理上点数据的形式，这样每个观测值确切或者近似的位置是一个连续的地质结构(topology)，而不是想象中的一个格子。地理统计学方法试图根据某个地理区域具体位置的信息创建空间共变模型，从而使连续的地理学变成地理统计形式。一种方法称为克里金(Kriging[16])法，由马特隆(Mtheron, 1963)正式发展而来，但以南非采矿工程师丹尼·克里格(Danie G. Krige)的名字命名，因为他开创了距离加权后测绘平均黄金等级的方法。这种方法广泛用于地球物理科学，现在也用于社会科学(Cho & Gimpel, 2007)。尽管以往研究用到的都是大样本汇总层面的数据或者没有空间识别特征的数据，但是现在可用的地理细分数据或者明显与地理相关的数据变得越来越多。

多层模型

在贝斯格(Besag, 1974)的早期贡献之后，有相当多的研究讨论了条件自回归模型(Conditionally Autoregressive Model, CAR)。在条件模型中，在某个地点观测到的随机变量取决于其邻近的外生观测值。在多变量和多层模型中，不

仅空间滞后变量需要是外生的，而且其他的解释变量也需要为外生。当前很多研究都在利用这种方法，有的是关注几个结果因变量。当前这些工作可以见金，班纳吉和卡林(Jin, Banerjee, and Carlin, 2007)以及茹和海尔德的研究(Rue and Held, 2005)。

另一种建立空间变化模型的相关方法是用多层方法考察本地变化的来源。多层空间模型是将不同分析层次的不确定性来源整合在一起。这种模型通过概率分布将不同层次的分析联系起来。在民主与发展的例子中，可以包括如下层次：(1)本国国内政党派系和机构的变化，它们会影响日常政治和经济的波动；(2)邻里效应，指某个相邻国家和另一国家之间存在很强的联系并且受其影响；(3)基于一系列国家所产生的区域效应，包括有些沿着地区边界运作的组织[①]；(4)全球化的力量，它会不同程度地影响到每一个国家，比如全球市场上的某些商品。模型中如果明确和详细说明了这些变化的来源，则属于多层模型。

当前处理这种观点的方法都是基于贝叶斯方法，也就是它依赖于迭代法[马尔可夫链蒙特卡罗方法(Markov chain Monte Carlo)、吉布斯抽样(Gibbs Sampling)、Metropolis-Hasting 等算法]来得到一套所有层次的空间过程的参数分布。这种方法需要很多计算，但是很有前景。当前的 R 软件包 spBayes 也可以帮助实现单变量和多变量空间模型的马尔可夫链蒙特卡罗计算(Finley, Banerjee & Carlin, 2007)。沃勒、卡林、夏和盖尔芬德(Waller, Carlin, Xia and Gelfand, 1997)提供了一项很有影响的应用，班纳吉等人(Banerjee et al., 2004)为多层方法提供了很好的综述。

时间序列数据

我们已经讨论了在同一时间段里横截面观测数据的模型估计。社会科学的很多分析都是基于时间序列的横截面(Time Series Cross Section, TSCS)数据结构，也就是同一个个体有多个不同时间点上的观测。空间滞后 y 模型也可

① 比如前文提到的经济合作和发展组织(Organization for Economic Cooperation and Development, OECD)就是这样一种情况。该组织中的成员国之间彼此的相互影响可能大于来自于其他国家的影响。——译者注

以扩展到 TSCS 数据的情况：

$$y_{i,t} = x_{i,t}\beta + \rho w_i y_{i,t} + \epsilon_{i,t}$$

该模型可能会面临时间上的序列相关问题，因为 $y_{i,t}$ 可能会和 $y_{i,t-1}$ 非常相似，进而造成误差独立假设的问题。一种解决方法是通过加入 y 的时间间隔，从而有：

$$y_{i,t} = \phi y_{i,t-1} + x_{i,t}\beta + \rho w_i y_{i,t} + \epsilon_{i,t}$$

如果我们想要同时说明时间和空间的依赖性，估计包含联立空间依赖性的 TSCS 模型是相当困难的。如果我们在右边加入滞后因变量，y 的误差项 $\in$ 的雅克比行列式转换(Jacobian of the transformation)将变得相当复杂，并且据我们所知，当前还没有人得到这种模型的满意估计。然而，如果假设 $y_{i,t}$ 对相邻的 y 的影响的发生具有一个时间滞后(比如 $y_{i,t-1}$)的话，就可以用 OLS 方法，因为与 y 相关的邻近值可以被视为在时间 t 之前就预先确定。这将得到：

$$y_{i,t} = \phi y_{i,t-1} + x_{i,t}\beta + \rho w_i y_{i,t-1} + \epsilon_{i,t}$$

空间效应中引入时间滞后量通常被认为和假设瞬时效应一样都是合理的。此外，还可以通过对模型估计残差进行适当的检验，尤其是交叉检验和样本外的探索方法，来检验模型在多大程度上成功解释了空间和时间依赖性(更多讨论，见 Beck et al.，2006)。

小结

空间依赖关系在很多社会现象中都发挥着重要作用。将空间层面引入分析中也是完全可行的，但是需要附加一些假设和信息。随着统计和计算机技术的发展，空间数据分析的障碍得以消减，我们期待它能给社会科学家们感兴趣的社会和空间过程带来新的见解。以往的经验让我们相信社会科学数据中具有很多未发现的依赖特征。只要将这些特征中的一部分考虑进来就可以产生全新的重要的启发。

附录 | 软件选项

在很长一段时间里，标准统计软件包都不能提供空间估计值，这使得感兴趣的研究者需要自己编程序或者购买安瑟林的 *SpaceStat* 软件。在过去几年中这种情况发生了很大的改变。在本部分中，我们将介绍一些可用的软件选择。

很多软件选项都依赖于奥德方法(Ord Approach)，在最优化之前估计矩阵 W 的特征值，不过现在一些其他方法也用更快的佩斯和巴里方法(Pace and Barry approach)。很多软件包都要求输入 $n \times n$ 的全秩矩阵。但这对于大样本的数据集来说很难做到。因为通常连接矩阵中会有很多 0 存在，软件选项中如果有用稀疏矩阵(sparse matrix)表示的话，就可以应用到更大的数据集上。

这里我们列出空间分析的一些软件选项。

(1) 安瑟林 *SpaceStat* 软件如今已不是安瑟林个人掌管的一个单独的软件包，而是卖给了一个商业公司，成为地理视觉化程序 *TerraSeer* 的一部分(见 http://www.terraseer.com/products_spacestat.php)。该软件的费用非常昂贵，即便是对学术用途而言也是。旧版本的 *SpaceStat* 在 MS-DOS 系统中运行，使用传统的菜单式界面。它也是依赖于奥德方法来计算加权矩阵的行列式，同时需要全部的矩阵表达式以进行估计。我们对当前的 *TerraSeer* 产品不太了解，因此不知它和以前的版本有何差别。

(2) 安瑟林和他的同事们开发出一种新的软件包叫做 GeoDa，可见 http://www.geoda.uiuc.edu/。GeoDa 可以进行探索性空间数据分析和简单的空间回归分析。GeoDa 界面完全是通过点击来完成的，而不需要任何编程；然而它不允许用户自定义或者修改其中的设置，这和一般化的统计软件包不同。安瑟林(Anselin, Syabri, Kho, 2004)等人表明这种软件主要是作为初学软件包，当

用户了解了 DeoDa 中的技术之后，可以逐步过渡到 R(p. 3)。

(3) 皮萨蒂(Pisati)编写的 *Stata* 宏 spatreg 可以估计空间自回归和误差模型。这个程序或者宏不好的地方在于它依赖于奥德方法，因而需要全部的矩阵表达式。*Stata* 的标准版本(standard Intercooled version)里同样还对矩阵大小有限制。参见 *Stata* 的技术报告(technical Bulletin)sg162。安装帮助见 *Stata* 中的 help stb。

(4) 比万德(Roger Bivand)开发了本书中讨论的模型的 R 软件包(*spdep*)。这个软件包还允许连接矩阵具有稀疏形式(sparse list)。比万德还开发了各种整合 R 和 GRASS 的材料，GRASS 是一种开源的 GIS 程序。此外，很多功能还可以用来制作地图，以及从 R 的 *Arcview* 的形文件(shapefiles)中提取信息。更多有关该软件包的细节可见：http://cran. r-project. org/src/contrib/Descriptions/spdep. html。这些功能和软件后台都是开放源代码并且免费的。

(5) 一些有关空间分析的 *MATLAB* 教材也已出版。佩斯和巴里(Pace and Barry)的空间统计学(Spatial Statistics)可以在 http://www. spatial-statistics. com/免费获得。MATLAB 本身并不是免费的。勒萨热(LeSage)的空间计量经济学(Spatial Econometrics)工具箱，可以在 http://www. spatial-econometrics. com 获取，它在估计大样本的数据集时非常有用，并且可以用于利用 saw()命令，估计具有两个连接矩阵的空间自回归模型。

(6) 来自 ESRI 公司最新版本的商业软件包 *ARCINFO* 包括很多可以用于空间形式数据集的统计分析工具，尤其是其中的 Statistical Analyst 工具包。它特别擅长于计算邻里数据和进行分类分析。

(7) Splus 来自于 Insightful 公司。和 R 一样，它也是基于 S 统计语言。它包括一个能提供很多空间相关数据分析的模块(SpatialStats)。其他工具还包括地理统计的、点状的以及格子状的空间数据。

(8) WINBUGS(http://www. mrc-bsu. cam. ac. uk/bugs/)和 GeoBUGS 是两个针对贝叶斯分析的程序。GeoBUGS 由一个流行病学家开发来作为 WinBUGS 的附加程序。它支持(相对较小的)空间模型的贝叶斯分析。

(9) 斯卡本伯格和果特威(Schabenberger and Gotway, 2005)提供了 SAS

中分析空间数据的宏和程序的扩展集，这些可以从出版商的网页上找到：www.crcpress.com。

（10）空间多层方法可以很容易在R软件包 spBayes 中找到，它可以进行常用的 MCMC(Markov Chain Monte Carlo)计算(Finley et al.，2007)。

注释

[1] 然而在社会科学中很少应用点数据。一个例外来自于最近曹和金佩尔(Cho and Gimpel, 2007)的研究。

[2] 在空间点过程条件下,有时这被称为合算统计量(Join-count statistics),因为它们计算了包含相似联合分布的邻近点的数量。

[3] 葛兰德(Grenander, 1954)指出即便对于均值的最小无偏估计也不应该忽略相关联的观测值的值。

$$\hat{\mu}=\frac{[y_1+(1-\rho)\sum_{i=2}^{n-1}y_i+y_n]}{[n-(n-2)\rho]}$$

[4] 吉尔里(Geary)和莫兰(Moran)都对很多领域做出了重要贡献。吉尔里因斯通—吉尔里效用函数(Stone-Geary utility function)和计算真实收入的等同购买力的国际比较而出名。

[5] 数据获取方式 http://privatewww.essex.ac.uk/~ksg/mindist.html。

[6] 莫兰 I 用于处理单个变量。当模型中明显存在多个变量时,比如 OLS 残差的例子,我们推荐使用一个略微修改的莫兰统计量,用于防止过高估计空间相关性。实际上,由于差别细微,通常在两种情况下都使用标准的莫兰统计量。特菲尔斯多夫(Tiefelsdorf, 1972)发展出一种潜在分布的鞍点(saddlepoint)近似法,这种方法在匹配分布尾部的很多情况下都比莫兰 I 要好。在这个问题上我们感谢罗杰·比万德(Roger Bivand)的周到建议。R 的 spdep 函数 lm.morantest.sad()可以执行特菲尔斯多夫的鞍点法。一个更直观的方法是利用拉格朗日乘子检验(Lagrange multiplier test)来检验空间自相关的具体形式,在下一章当中我们将会详细叙述。

[7] 正如海宁(Haining, 2003, pp. 276—283)提出,如果自变量之间存在很强的空间相关性,标准莫兰 I 可能过高估计残差的空间相关性。

[8] 用巴尔多夫—尼尔森(Barndorff-Nielsen)鞍点调整方法得到略微减小的估计:相对应的值为 6.9。

[9] 还可以设想出更多类似的分解方法(decomposition),比如引入多层协变量使得它们仅仅影响到某个具体区域或者行政区内的观测值。本书中将不讲解这种方法。

[10] 2004 年选举数据可以从 http://www.fec.gov/pubrec/fe2004/federalelections2004.pdf 获取。

[11] 有关非嵌套检验的例子可以参见 Clarke(2001)。

[12] 对该种以及其他依赖关系更详尽的考查可以参考 Ward and Hoff(2007);针对二元因变量,见 Ward, Siverson, and Cao(2007)。

[13] Wasserman and Faust(1994)总结过这种成对数据中三个一组的情况,主要是社会网络分析。

[14] 首位数字法则(first-digit law),以物理学家弗兰克·班佛命名,它说明数据中的第一位数字最常见为 1,越大的数字将越少见,或者更确切地说,数字出现的频数接近于。这种法则适用于大样本分布的、自然产生的数据,同时它还表明如果数据的分布和这种分布相差很大,则表明数据质量较低甚至可能是捏造的数据。

[15] 见 http://privatewww.essex.ac.uk/~ksg/polity-data.html。

[16] 根据发明者的名字,Kriging 应当被读作"Kricking"。

参考文献

Adolph, C. A. (2004). *The dilemma of discretion: Career ambitions and the politics ofcentral banking*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht, The Netherlands: Kluwer.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical Analysis*, *27*, 93—115.

Anselin, L., Syabri, I., & Kho, Y. (2004). *GeoDa: An introduction to spatial data analysis (Typescript)*. Urbana-Champaign: Department of Agricultural and Consumer Economics, University of Illinois.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: Chapman & Hall.

Baybeck, B., & Huckfeldt, R. (2002). Urban contexts, spatially dispersed networks, and the diffusion of political information. *Political Geography*, *21*, 195—220.

Beck, N., Gleditsch, K. S., & Beardsley, K. (2006). Space is more than geography: Using spatial econometrics in the study of political economy. *International Studies Quarterly*, *50*, 27—44.

Beck, N., & Katz, J. N. (1996). Nuisance vs. substance: Specifying and estimating timeseries—Cross-section models. *Political Analysis*, *6*, 1—36.

Berk, R. A., Western, B., & Weiss, R. E. (1995). Statistical inference for apparent populations (with discussion). *Sociological Methodology*, *25*, 421—485.

Besag, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society, Series B, Methodological*, *34*, 75—83.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, *36*, 192—225.

Bivand, R. (2002). Spatial econometrics functions in R: Classes and methods. *Journal of Geographical Systems*, 4, 405—421.

Bivand, R., Pebesma, E., & Gomez-Rubio, V. (forthcoming). *Applied spatial data analysis with R*. New York: Springer.

Boots, B. N., & Tiefelsdorf, M. (2000). Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems*, *2*, *319*—348.

Brundson, C., Fotheringham, A. S., & Charlton, M. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, *28*, 281—298.

Burkhart, R., & Lewis-Beck, M. (1994). Comparative democracy: The economic development thesis. *American Political Science Review*, *88*, 903—910.

Calvo, E., & Escolar, M. (2003). The local voter: A geographically weighted approach to ecological inference. *American Journal of Political Science*, *47*, 189—204.

Cho, W. K. T., & Gimpel, J. G. (2007). Prospecting for (campaign) gold. *American Journal of Political Science*, *51*, 255—268.

Christensen, O. F. , & Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, *58*, 280—286.

Clarke, K. A. (2001). Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science*, *45*, 724—744.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Cliff, A. D. , & Ord, J. K. (1971). Evaluating the percentage points of a spatial all tocorrelation coefficient. *Geographical Analysis*, *4*, 51—62.

Cressie, N. A. C. (1993). *Statistics for spatial data* (rev. ed.). New York: Wiley.

Dalgaard, P. (2002). *Introductory statistics with R*. Berlin: Springer.

Deutsch, K. W. , & Isard, W. (1961). A note on a generalized concept of effective distance. *Behavioral Science*, *6*, 308—311.

Feenstra, R. C. , Rose, A. K. , & Markusen, J. R. (2001). Using the gravity model to differentiate among alternative theories of trade. *Canadian Journal of Economics*, *34*, 430—447.

Finley, A. O. , Banerjee, S. , & Carlin, B. P. (2007, April). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, *19*(4). Retrieved October 29, 2007, from http://www.jstatsoft.org/v19.

Fotheringham, A. S. , Charlton, M. , & Brundson, C. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. New York: Wiley.

Franzese. R. (1999). Partially independent central banks, politically responsive governments, and inflation. *American Journal of Political Science*, *43*, 681—706.

Franzese, R. , & Hayes, J. C. (2007). Spatial econometric models for the analysis of TSCS data in political science. *Political Analysis*, *15*, 140—164.

Gartzke, E. (1998). Kant we all just get along? Opportunity, willingness and the origins of the democratic peace. *American Journal of Political Science*, *42*, 1—27.

Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, *61*, 101—107.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*, 320—328.

Getis, A. , & Boots, B. (1978). *Models of spatial processes*. Cambridge, UK: Cambridge University Press.

Getis, A. , & Ord, J. K. (1996). Local spatial statistics: An overview. In P. Longley & M. Batty (Eds.), *Spatial analysis: Modelling in a GIS environment* (pp. 261—277). Cambridge, UK: Geoinformation International.

Geyer, C. J. , & Thompson, E. A. (1992). Constrained Monte Carlo, maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *Methodological*, *54*, 657—699.

Gleditsch, K. S. (2002a). *All international politics is local: The diffusion of conflict, integration, and democratization*. Ann Arbor: University of Michigan Press.

Gleditsch, K. S. (2002b). Expanded trade and GDP data. *Journal of Conflict Resolution*, *46*, 712—724.

Gleditsch, K. S. , & Ward, M. D. (1997). Double take: A re-examination of democracy and autoc-

racy in modern polities. *Journal of Conflict Resolution*, *41*, 361—382.

Gleditsch, K. S., & Ward, M. D. (2000). War and peace in time and space: The role of democratization. *International Studies Quarterly*, *44*, 1—29.

Gleditsch, K. S., & Ward, M. D. (2001). Measuring space: A minimum distance database and applications to international studies. *Journal of Peace Research*, *38*, 749—768.

Gleditsch, K. S., & Ward, M. D. (2007). The diffusion of democracy and the international context of democratization. *International Organization*, *60*, 911—933.

Grenander, U. (1954). On the estimation of regression coefficients in the case of autocorrelated disturbance. *Annals of Mathematical Statistics*, *25*, 252—272.

Griffith, D. A. (1996). Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In S. Arlinghaus (Ed.), *Practical handbook of spatial statistics* (pp. 65—83). Boca Raton, FL: CRC Press.

Griffith, D. A. (2003). Using estimated missing spatial data with the 2-median model. *Annals of Operations Research*, *122*, 233—247.

Haining, R. (2003). *Spatial data analysis: Theory and practice* (1st ed.). Cambridge, UK: Cambridge University Press.

Holmes, T. J. (2006, February). *Geographic spillover and unionism*. National Bureau of Economic Research (Working Paper Series 12025). Retrieved October 17, 2007, from http://www.nber.org/papers/w 12025.

Hubert, L. J., Golledge, R. G., & Constanzo, C. M. (1981). Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, *12*, 224—233.

Huffer, F. W., & Wu, H. (1998). Markov chain Monte Carlo for autologistic, regression models with application to the distribution of plant species. *Biometrics*, *54*, 509.

Imai, K. (2005). Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review*, *99*, 283—300.

Işik, O., & Pinarcioğlu, M. M. (2007). Geographies of a silent transition: A geographically weighted regression approach to regional fertility differences in Turkey. *European Journal of Population*, *22*, 399—421.

Jaggers, K., & Gurr, T. R. (1995). Tracking democracy's "Third Wave" with the Polity Ⅲ data. *Journal of Peace Research*, *32*, 469—482.

Jin, X., Banerjee, S., & Carlin, B. P. (2007). Order-free coregionalized areal data models with application to multiple disease mapping. *Journal of the Royal Statistical Society*, *Series B*, *69*, 817—838.

Johnson, S. (2006). *The ghost map*. New York: Riverhead Books.

Jones, D. M., Bremer, S. A., & Singer, J. D. (1996). Militarized interstate disputes, 1816—1992: Rationale, coding rules, and empirical applications. *Conflict Management and Peace Science*, *15*, 163—213.

Keele, L., & Kelly, N. J. (2006). Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political Analysis*, *14*, 186—205.

Kenny, D. (1981). Interpersonal perception: A multivariate round robin analysis. In M. B. Brewer

& B. E. Collins (Eds.), *Scientific inquiry and the social sciences: A volume in honor of Donald T. Campbell* (pp. 288—309). San Francisco: Jossey-Bass.

Kidron, M. (1981). *The state of the world atlas*. New York: Simon & Schuster.

Lacombe, D. (2004). Does econometric methodology matter? An analysis of public policy using spatial econometric techniques. *Geographical Analysis*, *36*, 105—118.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with non-experimental data*. New York: Wiley.

Lee, C.-S. (2005). Income inequality, democracy, and public sector size. *American Sociological Review*, *70*, 158—181.

Leontief, W. W. (1986). *Input-output economics*. New York: Oxford University Press.

Lin, T.-M., Wu, C.-E., & Lee, F. Y. (2006). Neighborhood influence on the formation of national identity in Taiwan: Spatial regression with disjoint neighborhoods. *Political Research Quarterly*, *59*, 35—46.

Lipset, S. M. (1959). Some social requisites of democracy. *American Political Science Review*, *53*, 69—105.

Lofdahl, C. (2002). *Environmental impacts of globalization and trade: A systems study*. Cambridge: MIT Press.

Malloy, T. E., & Kenny, D. A. (1986). The social relations model: An integrative method for personality research. *Journal of Personality*, 54, 199—225.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*, 1246—1266.

Moran, P. A. P. (1950a). Notes on continuous stochastic phenomena. *Biometrika*, *37*, 17—23.

Moran, P. A. P. (1950b). A test for serial independence of residuals. *Biometrika*, *37*, 178—181.

Morrow, J. D., Siverson, R. M., & Tabares, T. E. (1998). The political determinants of international trade: The major powers, 1907—90. *American Political Science Review*, *92*, 649—661.

Murdoch, J. C., Sandler, T., & Sargent, K. (1997). A tale of two collectives: Sulfur versus nitrogen oxides emission reduction in Europe. *Economics*, *64*, 281—301.

Ord, J. K. (1975). Estimation methods for models of spatial interactions. *Journal of the American Statistical Association*, *70*, 120—126.

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, *27*, 286—306.

Pollins, B. M. (1989a). Conflict, cooperation, and commerce: The effect of international political interactions on bilateral trade flows. *American Journal of Political Science*, *33*, 737—761.

Pollins, B. M. (1989b). Does trade still follow the flag? A model of international diplomacy and commerce. *American Political Science Review*, *83*, 465—480.

R Development Core Team. (2004). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing (ISBN 3-900051-00-3; http://www.R-project.org).

Ripley, B. D. (1981). *Spatial statistics*. New York: Wiley.

Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge, UK: Cambridge Uni-

versity Press.

Rose, A. K. (2004). Does the WTO really increase trade? *American Economic Review*, *94*, 98—114.

Rozanski, J., & Yeats, A. (1994). On the (in)accuracy of economic observations: An assessment of trends in the reliability of international trade statistics. *Journal of Development Economics*, *44*, 103—130.

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. London: Chapman & Hall.

Schabenberger, O., & Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. Boca Raton, FL: Chapman & Hall.

Shin, M. E. (2001). The politicization of place in Italy. *Political Geography*, *20*, 331—352.

Shin, M. E., & Agnew, J. (2002). The geography of party replacement in Italy, 1987—1996. *Political Geography*, *21*, 221—242.

Shin, M. E., & Agnew, J. (2007a). *Berlusconi's Italy: Where it started, where it ended*. Philadelphia: Temple University Press.

Shin, M. E., & Agnew, J. (2007b). The geographical dynamics of Italian electoral change, 1987—2001. *Electoral Studies*, *26*, 287—302.

Signorino, C., & Ritter, J. (1999). Tau-b or not tau-b. *International Studies Quarterly*, *43*, 115—144.

Tiefelsdorf, M. (1972). The saddlepoint approximation of Moran's I and local Moran's I: Reference distributions and their numerical evaluation. *Geographical Analysis*, *34*, 187—206.

Tufte, E. R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1992). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

Varian, H. R. (1972). Benford's law., *American Statistician*, *26*, 65.

Wainer, H. (2004). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton, NJ: Princeton University Press.

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, *121*, 311—324.

Waller, L. A., Carlin, B. P., Xia, H., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, *92*, 607—617.

Ward, M. D., & Gleditsch, K. S. (2002). Location, location, location: An MCMC approach to modeling the spatial context of war and peace. *Political Analysis*, *10*, 244—260.

Ward, M. D., & Hoff, P. D. (2007). Persistent patterns of international commerce. *Journal of Peace Research*, *44*, 157—175.

Ward, M. D., Siverson, R. M., & Cao, X. (2007). Disputes, democracies, and dependencies: A reexamination of the Kantian peace. *American Journal of Political Science*, *51*, 583—601.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cam-

bridge, UK: Cambridge University Press.

Watts, D. J. (2003). *Six degrees: The science of a connected age*. New York: W. W. Norton.

West, W. J. (2005). Regional cleavages in Turkish politics: An electoral geography of the 1999 and 2000 national elections. *Political Geography*, *24*, 499—523.

You, J.-S., & Khagram, S. (2005). A comparative study of inequality and corruption. *American Sociological Review*, *70*, 136—157.

译名对照表

ad hoc	事先设定的
apparent population	明显的总体
areal data	地区数据
autologistic	自相关逻辑斯蒂
Benford's Law	班佛定律
bias	偏误
Collegi	学院
Conditionally autoregressive model，CAR	条件自回归模型
cross-validation	交叉验证
distribution of first digits	首位数字分布
equilibrium impact	均衡效应
fixed effects	固定效应
Geary's C	吉尔里 C 统计量
Gibbs Sampling	吉布斯抽样
global correlation	全球相关
Imputation	插补方法
inconsistency	不一致
Intercept	截距
inverse	逆矩阵
Jacobian of the transformation	雅克比行列式转换
kernel density	核密度分布
lattice data	晶格数据
Local Indicator of Spatial Association，LISA	局部空间相关指标
map mashups	地图混搭程序
Markov chain Monte Carlo	马尔可夫链蒙特卡罗方法
Maximum likelihood estimator，MLE	最大似然估计量
Mean Squared Errors	均方误
mean-normalized cross-product	均值正态化后得到的内积
metric	度量学
Monte Carlo simulation	蒙特卡洛模拟
Moran I	莫兰 I 统计量
Ord Approach	奥德方法
out-of-sample prediction test	样本外的预测检验
Parsimony	简约性
point data	点状数据

point processes	点过程
pooled OLS	合并数据的 OLS
rug plot	毯图
Shin Plot	锡恩图
Shin spatial scatterplot	锡恩空间散点图
simultaneity	共时性
simultaneous model	联立模型
sparse list	稀疏形式
sparse matrix	稀疏矩阵
spatial autoregressive	空间自回归
Spatial Econometrics	空间计量经济学
Spatial Error	空间性误差
Spatial lag	空间滞后
Spatial Statistics	空间统计学
Spatially Lagged Dependent Variable	空间滞后因变量
super-population	超总体
Time Series Cross Section，TSCS	时间序列的横截面